Logic-based Analysis of Gene Expression Data Predicts Pathway Crosstalk between TNF, TGFB1 and EGF in Basal-like Breast Cancer

Kyuri Jo[1,+], Beatriz Santos Buitrago[1,+], Minsu Kim[2], Sungmin Rhee[1], Carolyn Talcott[3], and Sun Kim[1,2,4]

[1]Department of Computer Science and Engineering, Seoul National University, Seoul, 08826, Korea.
[2]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, 08826, Korea.
[3]SRI International, Menlo Park, CA 94025, USA.
[4]Bioinformatics Institute, Seoul National University, Seoul, 08826, Korea.
[+]Equal contributors
[*]To whom correspondence should be addressed (sunkim.bioinfo@snu.ac.kr)

April 20, 2019

Abstract

For breast cancer, clinically important subtypes are well characterised at the molecular level in terms of gene expression profiles. In addition, signaling pathways in breast cancer have been extensively studied as therapeutic targets due to their roles in tumor growth and metastasis. However, it is challenging to put signaling pathways and gene expression profiles together to characterise biological mechanisms of breast cancer subtypes since many signaling events result from post-translational modifications, rather than gene expression differences.

We present a logic-based approach to explain the differences in gene expression profiles among breast cancer subtypes using Pathway Logic and transcriptional network information. Pathway Logic is a rewriting-logic-based formal system for modeling biological pathways including post-translational modifications. Proposed method demonstrated its utility by constructing subtype-specific path from key receptors (TNFR, TGFBR1 and EGFR) to key transcription factor (TF) regulators (RELA, ATF2, SMAD3 and ELK1) and identifying potential pathway crosstalk via TFs in basal-specific paths, which could provide a novel insight on aggressive breast cancer subtypes.

Availability: Analysis result is available at http://epigenomics.snu.ac.kr/PL/

1 Introduction

Breast cancer is the first cause of cancer death in women with a worldwide mortality rate of 12.9% [1]. A better understanding of the pathogenesis of the disease is a key to improving current therapies and discovering new treatment options. Especially, signaling pathways in breast cancer such as tumor necrosis factor (TNF), transforming growth factor beta 1 (TGFB1) and epidermal growth factor (EGF) have been studied as therapeutic targets due to their roles in tumor growth and metastasis [2, 3, 4]. On the other hand, genome-wide gene expression profiles are measured for many breast cancer patients using gene expression measurement technologies such as gene expression microarray and sequencing machines. As a result, clinically important breast subtypes are well characterised at the molecular level in terms of gene expression profiles [5, 6]. However, there is little study in combining signaling pathways and gene expression profiles together to characterise biological mechanisms in each of breast cancer subtypes since many signaling events result from post-translational modifications (PTMs) followed by gene expression differences [7], which is difficult to be identified using the existing pathway models. The goal of this study is to provide an integrated analysis of signaling pathways and gene expression profiles to determine subtype-specific path from key receptors to key transcription factor (TF) regulators to downstream genes targeted by TF. There are several challenges for achieving this research goal.

One approach to model cancer phenotype is to construct *in silico* gene regulatory networks (GRNs) based on the putative relationship (e.g. activation or inhibition) between genes inferred from gene expression data. Network inference algorithms like ARACNE [8] have been used to elucidate the underlying mechanisms of breast cancer such as Epithelial-Mesenchymal Transition (EMT) [9] and metastasis [10]. As breast cancer subtypes can be defined at the molecular level using gene expression information [5], these network-based machine learning approaches are very powerful to bring novel insight into breast cancer biological mechanisms. However, *in silico* network generated from gene expression without prior knowledge requires additional analysis steps due to a number of false positive relationships.

To overcome the limitation of gene expression data, several algorithms have incorporated the information of biological pathways into the analysis in addition to gene expression data [11]. Such pathway analysis algorithms use networks generated from the curated knowledge of biological pathways provided by public databases such as KEGG [12] and NCI PID [13], rather than the network generated by *in silico* prediction. Thus, they charts a more precise and compact landscape of dysregulated biological processes, suitable for comparing networks from multiple phenotypes. When converting the information from the public databases into a network, however, most of the pathway analysis tools such as Clipper [14] construct a simple gene-gene network without PTM reactions (e.g. phosphorylation or ubiquitination), leading to a loss of information on pathways [15]. Simplified network provides a limited view of cancer subtypes since many biological events such as signal transduction can occur without significant changes in expression levels. For instance, a signaling pathway includes a sequence of PTM reactions that can modify the activity of TF proteins prior to their transcriptional regulation on the target genes [16].

Logic modeling is another approach to model and analyse a complex pathway. In such models, molecular events including PTMs are defined as general rules and a set of rules makes up a knowledge base. By applying rules to the premise (initial state of the cell), the user can derive a conclusion (final state of the cell). Several logic-based models have been developed so far [17], especially boolean models with two binary states for floral morphogenesis [18], mammallian cell cycle [19], apoptosis [20] and hepatocyte growth factor (HGF) [21]. Recently, BioASF [22] suggested a general framework for pathway models specified in BioPAX language [23] that has been used by several pathway databases [24, 25]. Although these models delineate each biological process in more detail than a gene-gene network used in the pathway analysis, the main purpose of the logic models is the simulation study to enumerate possible final states or concentrations of the product, not designed for integrating gene expression data and performing comparative studies between phenotypes.

Pathway Logic (PL) [26] is a general system for modeling signaling and metabolic pathways in Maude language [27]. PL provides several knowledge bases including STM7 knowledge base with 32 models of signal transduction for different stimuli. Models in PL are constructed with detailed descriptions of biological process including cellular locations, interactions and PTMs of the molecules. Unlike the other logic models, PL system provides internally implemented functions such as comparison between two different models and dynamic searching of subgraph with certain endpoint conditions, showing a wide range of application for the pathway analysis beyond the typical simulation studies. Recently, phosphoproteomics data has been integrated into the PL system to derive different response of signaling pathways to different drugs in SKMEL133 melanoma cancer cells [28]. This research work shows that integrating the PL system and omics data can effectively narrow the range of rules in deductive reasoning by specifying the condition-specific cell states observed in the experiments. In fact, gene expression data has much more potential for the integrated system as millions of omics data sets have been generated from high-throughput experiments over the past decade [29].

A possible limitation of the pathway analysis might be the small set of genes belonging to one pathway (e.g. 29 to 351 genes in KEGG signaling pathways), not enough to analyse the global changes triggered by the signal transduction. One of the methods to expand the pathway analysis of a limited set of genes is to incorporate TFs. TFs are principal regulators of gene expression in eukaryotic cells [30, 31]. TFs regulate their target genes to mobilize the appropriate protein response and thereby contribute to a global change in cellular states, even resulting in a different phenotype of an organism [32, 31]. Most of the signaling pathways described in the database start from a ligand binding to a receptor and the transmitted signal activates a number of proteins, including TFs. Therefore, TFs can work as a bridge from signal transduction to broader changes in a cell or an organism.

In this study, we propose a novel computational framework to integrate pathway models of PL, gene expression data and transcriptional network to explore subtype-specific biological mechanisms in breast cancer. By rewriting rules in PL models, we derived paths of signal transduction from key receptors to key TF regulators and to downstream genes targeted by TFs. TFs included in the signaling paths are statistically tested based on their contribution to the subtype-specific expression of their target genes. Key features of the proposed method compared to existing methods are as follows:

1. Pathway models defined in PL include detailed information on biological entities such as PTMs that are often ignored in *in silico* network or current pathway network structure with limited interaction types.

2. Logical inference-based functions implemented in PL can be used for pathway analysis such as searching for sub-pathways satisfying a specific condition and also for comparing multiple pathways, which are not available in the existing logic models designed for simulation studies.

3. Integration of gene expression data with PL models and PL functions provide a more accurate and comprehensive landscape of signaling pathways —e.g. signaling paths activated only in a certain phenotype (subtype) including the upstream PTM reactions.

We used our analysis framework to analyse The Cancer Genome Atlas (TCGA) breast cancer data [33] and determined subtype-specific signaling path from three receptors (TNFR, TGFBR1, EGFR) to four TFs (RELA, ATF2, SMAD3, ELK1) and identified their roles in breast cancer. Performance of our method is compared with a reverse engineering algorithm, ARACNE [8] and a pathway analysis algorithm, Clipper [14].

2 Methods

The computational framework proposed in this paper is based on the pathway models in PL. In PL models, steps of a signal transduction pathway are represented as occurrences and transition rules. An **occurrence** is a protein together with its state and location. Transition **rules**, representing biochemical reactions that change a state and/or location of the occurrences, are curated from the literature (**datum**) and collected into a knowledge base. In the context of such a knowledge base, a model is defined by specifying the set of occurrences initially present (called an initial state or **dish**). The graphical representation of occurrences, rules and dishes is shown in Figure 1 (a).

Among the knowledge bases in PL, STM7 is a model of intracellular signal transduction that responses to peptides, chemicals, or stresses added to cells. STM7 includes 32 dishes (initial states), each representing one stimuli. To analyse breast cancer dataset, we selected three dishes from STM7 knowledge base according to the relevance to cancer mechanism —response to Tumor Necrosis Factor (TNF), Transforming Growth Factor Beta 1 (TGFB1) and Epidermal Growth Factor (EGF) stimulation. TNF, TGFB and EGF signaling pathways are known to have strong relationship through pathway crosstalk in breast cancer cell line modulating the tumor phenotype [34]. Table 1 shows the brief description of these PL models. Initial states and representative rewrite rules (receptor binding event of each ligand) of three pathways are described in Supplementary Information.

2.1 Finding transcription factors mediating subtype-specific control of pathway crosstalk

It is well known that pathway crosstalk frequently occurs between signaling pathways to orchestrate cooperative response to various environmental stress [35]. The way two pathways communicate determines how cancer tissue reacts to oncogenic environment, which affects its potential malignancy. For example, Dent et al.

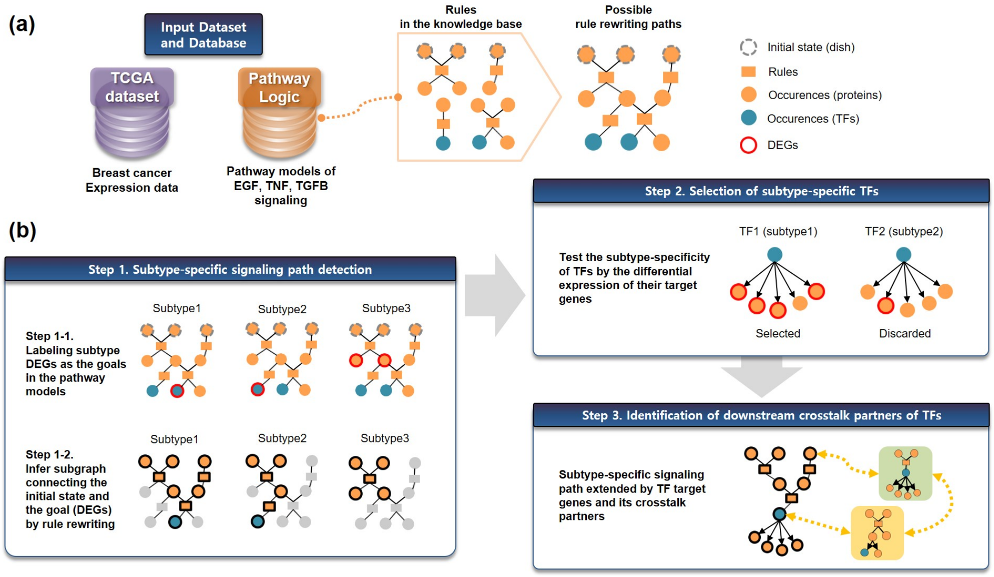
(a)
Input Dataset and Database
TCGA dataset
Pathway Logic
Breast cancer Expression data
Pathway models of EGF, TNF, TGFB signaling
Rules in the knowledge base
Possible rule rewriting paths
Initial state (dish)
Rules
Occurences (proteins)
Occurences (TFs)
DEGs
(b)
Step 1. Subtype-specific signaling path detection
Subtype1
Subtype2
Subtype3
Step 1-1. Labeling subtype DEGs as the goals in the pathway models
Subtype1
Subtype2
Subtype3
Step 1-2. Infer subgraph connecting the initial state and the goal (DEGs) by rule rewriting
Step 2. Selection of subtype-specific TFs
Test the subtype-specificity of TFs by the differential expression of their target genes
TF1 (subtype1)
TF2 (subtype2)
Selected
Discarded
Step 3. Identification of downstream crosstalk partners of TFs
Subtype-specific signaling path extended by TF target genes and its crosstalk partners

Figure 1: Workflow of the proposed method. (a) Two input dataset (database) for the analysis. Gene expression data was obtained from TCGA data portal [33] to analyze breast cancer subtypes. We used PL database to model signaling pathways, where molecular and cellular processes are represented with occurrences (orange circles) and rules (orange boxes). Some of the proteins in the occurrences are TFs (blue circles). (b) Three-step procedure for the analysis. To detect the subtype-specific signaling paths in breast cancer, differentially expressed genes (DEGs) were determined for each subtype and set as the goals of the rule rewriting (Step 1-1). By performing rewriting of rules in the model, we inferred subgraph of the signaling pathway starting from the initial state leading to DEGs (Step 1-2). The subgraphs include TFs, the subtype-specificity of which was tested by the differential expression of their target genes (Step 2). As a result, subtype-specific signaling paths were extended to the downstream target genes of TFs and their crosstalk partners.

revealed that a combination of pathways, including AKT-ERK crosstalk, affected cancer cell viability and had bad influence on prognosis [36].

TFs have been considered as important participants for regulating pathway crosstalk and also potential molecular targets. Blokzijl et al. suggested that TF can act as a mediator between different signaling pathways [37]. Such crosstalk between signaling pathways can be targeted by a molecular therapy if we can understand the underlying mechanism and specify targets [38]. The proposed algorithm can contribute in this regard since it finds TFs having potential as pathway crosstalk mediators and shows how the crosstalk is regulated. The analysis starts with signaling upstream factors (e.g. TNF, TGFB1, and EGF) and finds downstream paths leading to the pathway crosstalk mediated by a TF. This process can be summarized as three steps as described in Figure 1 (b).

- Step 1: Detecting subtype-specific paths from signaling upstream factors to potential TF mediators.
- Step 2: Selecting TFs that regulate subtype-specific expression of target genes.
- Step 3: Identifying downstream crosstalk partners mediated by TFs.

2.2 Subtype-specific signaling path detection by rule rewriting

Given that differentially expressed genes (DEGs) for each subtype of cancer are determined by statistical algorithms such as Limma [39], subtype-specific signaling path is defined as the path where the initial state of the pathway is rewritten to the state with activated subtype DEGs. Therefore, subtype DEGs up-regulated in cancer cells (tagged with `-act` modifier in PL models) are used as goals of the signaling in the PL

Table 1: Description of EGF, TNF and TGFB1 models in PL.

Dish	Ligand(s)	Receptor(s)	# of Rules	# of Datums
TnfDish	Tnf	TnfR1	87	1,915
Tgfb1Dish	Tgfb1	TgfbR1, TgfbR2	57	968
EgfDish	Egf	EgfR	129	2,281

model (Step 1-1 in Figure 1 (b))). For the four breast cancer subtypes, we had different sets of DEGs in comparison with the same normal samples (normal vs. luminal A, luminal B, basal-like or HER2-enriched) deriving different subtype-specific signaling paths.

In PL models, finding a path from an initial state to the goals (endpoints) can be performed in two ways: using the Maude language and the Pathway Logic Assistant (PLA). The Maude language provides the `search` command (Listing 1) that takes parameters (`[n]`: the first n solutions; `=>+`: at least one step) and target state of the endpoint defined by the operator `PD`. The argument of `PD` is the locations with their respective contents. Listing 1 shows an example of the search command to determine a path from EgfDish to the activated Erks protein. The contents of each location (e.g. `EgfRC`) are elements or variables (e.g. `thEgfRC:Things`). In this example, the target state of the endpoint is a protein `prot:BProtein` activated (`-act`) in the nucleus (`NUc`) with other modifications `mod:ModSet` and the variable `prot:BProtein` should belong to the `ErkS` protein family. The solution to this query is the matching state of the endpoint (Listing 2). Note that the variable `prot:BProtein` is matched to `Erks` and has modification of `phos (TEY)` and `phos (SPS)` in addition to activation. Asking Maude to `show path label [state ID]`, we can summarize rewrite rules applied to reach the state [40].

Listing 1: Example of search command in PL using Maude language. The code is to find a path from EgfDish to activation of Erks.

```
search [1] in QQ : EgfDish =>+ PD(
locOther:Locations {EgfRC | thEgfRC:Things [EgfR - Yphos] : Egf}
{XOut | thXOut:Things}
{CLm | thCLm:Things}
{CLi | thCLi:Things}
{CLc | thCLc:Things}
{NUc | thNUc:Things [prot:BProtein - mod:ModSet act]})
such that (prot:BProtein :: ErkS) = true .
```

Listing 2: Example of solution (matching state) for search command in PL. Erks is activated and phosphorylated at two different sites.

```
Solution 1 (state 17589)
locOther:Locations --> {CLo | empty}
thCLm:Things --> Erbb2 Pag1 Plscr1
thCLi:Things --> Gnai1 Gnai3 Pld1 [Cdc42 - GDP] [Hras - GTP] [Kras - GDP] [Nras - GDP] [Rac1 - GTP] [Rala - GDP] [Rit1 - GDP] Pld2
thXOut:Things --> empty
thCLc:Things --> Abl1 Akt1 Araf Cbl Cblb Crk CrkL Csk Dbl Erk5 Fak1 Gab1 Hpk1 IqGap1 Jak1 Jak2 Jnks Lkb1 Mek5 Mekk1 Mekk2 Mekk3 Nckipsd P38s
      Pdpk1 Pkcd Raf1 RapGef1 Rsk1 (Mlst8 : Mtor : Raptor) [Gsk3s - act] [Braf - act] [Mlk3 - act] [Mek1 - act] Shoc2 Sin1 Smad3 Stat1 Stat3
thNUc:Things --> Atf1 Creb1 Elk1 Fos HistH3 Ipo7 Jun Msk1 Msk2 Myc
prot:BProtein --> Erks
mod:ModSet --> phos (TEY) phos (SPS)
thEgfRC:Things --> Grb2 Pi3k [Fak2 - act] [Gab2 - Yphos] [Git1 - Yphos] [Pxn - Yphos] [Shp2 - Yphos] [Ptk6 - act] [Shc1 - Yphos] [Sos1 - Yphos
      ] [Src - act]
```

Another approach to derive subgraph in PL is the PLA. PLA can be executed online (http://pl.csl.sri.com/) and provides an interactive graphical interface to browse and query PL network models. Using PLA, we define endpoints as checking the 'make occ as goal' box for the target occurrences and created the subgraph by 'Subnet' button. In this study, we explored three dishes —TNF, TGFB1 and EGF where the goals were occurrences of DEGs in each breast cancer subtype and computed subgraphs that led to the activation of those genes. As a result, we obtained four different subgraphs for each dish, each subgraph for each breast cancer subtype (Step 1-2 in Figure 1 (b)).

2.3 Selection of subtype-specific transcription factors

Each subtype-specific signaling path may include TFs that may cause a global gene expression change in the cancer subtype samples. To ensure the significance of TFs as subtype-specific regulators, we test the differential expression of their target genes 'between subtypes' (e.g. basal-like vs. three other subtypes of breast cancer). We assume that if a transcription factor is a subtype-specific regulator, a ratio of differentially expressed target genes to all target genes should be much higher than a ratio of all DEGs to all genes in total. A hypergeometric test was used to measure the influence of TF comparing a ratio of total DEGs (n) to total genes (N) with a ratio of DEGs among the target genes (k) to total target genes of the TF (K) (Eq. 1). P-value was calculated as one minus cumulative distribution function of Eq. 1. The hypergeometric test is a simple and fundamental metric that has been used for finding pathways with a high proportion of DEGs [41].

$$P(X = k) = f(k; N, K, n) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{1}$$

Interaction information between TFs and their target genes in a GRN was from HTRIdb (The Human Transcriptional Regulation Interactions database) [42] that provides experimentally verified or computationally predicted TF-DNA-binding sites from six public databases and literature [43]. Additionally, we used MSigDB (Molecular Signatures Database) [44] where C3-TFT gene set (gene set of transcription factor targets) is provided [45].

2.4 Identification of the downstream crosstalk partners of TFs

For each TF detected as subtype-specific in Step 2, we identify the downstream crosstalk partners using their target genes. Among the target genes, we filter subtype DEGs that are up/down-regulated in the same subtype where TF is activated and their enrichment on biological pathways is tested using Enrichr [46] web service available at http://amp.pharm.mssm.edu/Enrichr/. We assume that pathways enriched with TF target genes are potential crosstalk partners of TFs, deriving a subtype-specific signaling circuit connecting the upstream pathway of TF and downstream crosstalk pathways.

3 Results

To identify the subtype-specific signaling paths and their crosstalk in breast cancer, RNA-seq data was downloaded from TCGA data portal [33]. Differentially expressed genes (DEGs) of four breast cancer

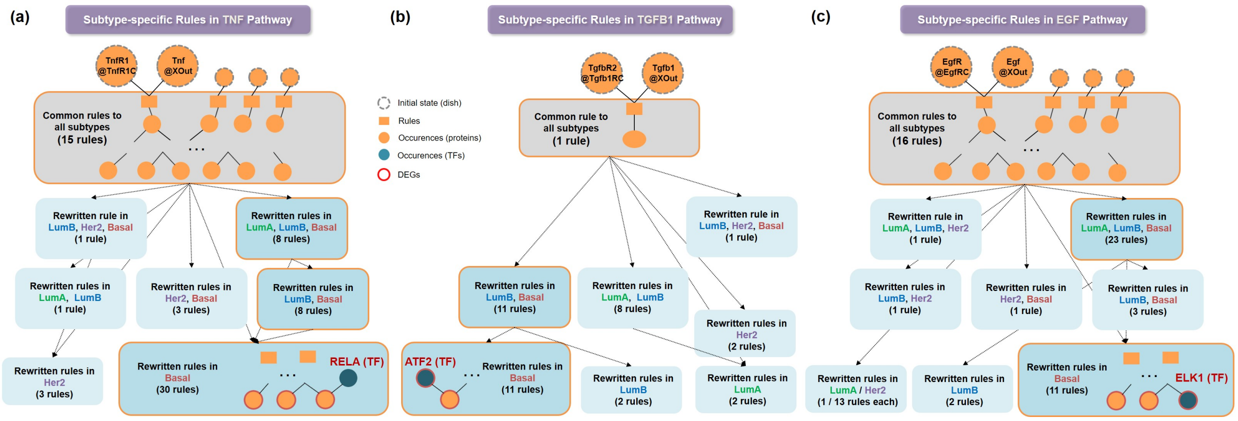
(a)
Subtype-specific Rules in TNF Pathway
TnfR1 @TnfR1C
Tnf @XOut
Common rules to all subtypes (15 rules)
...
Rewritten rule in LumB, Her2, Basal (1 rule)
Rewritten rules in LumA, LumB, Basal (8 rules)
Rewritten rules in LumA, LumB (1 rule)
Rewritten rules in Her2, Basal (3 rules)
Rewritten rules in LumB, Basal (8 rules)
Rewritten rules in Her2 (3 rules)
Rewritten rules in Basal (30 rules)
RELA (TF)
(b)
Subtype-specific Rules in TGFB1 Pathway
Initial state (dish)
Rules
Occurences (proteins)
Occurences (TFs)
DEGs
TgfbR2 @Tgfb1RC
Tgfb1 @XOut
Common rule to all subtypes (1 rule)
Rewritten rule in LumB, Her2, Basal (1 rule)
Rewritten rules in LumB, Basal (11 rules)
Rewritten rules in LumA, LumB (8 rules)
Rewritten rules in Her2 (2 rules)
ATF2 (TF)
Rewritten rules in Basal (11 rules)
Rewritten rules in LumB (2 rules)
Rewritten rules in LumA (2 rules)
(c)
Subtype-specific Rules in EGF Pathway
EgfR @EgfRC
Egf @XOut
Common rules to all subtypes (16 rules)
Rewritten rule in LumA, LumB, Her2 (1 rule)
Rewritten rules in LumA, LumB, Basal (23 rules)
Rewritten rules in LumB, Her2 (1 rule)
Rewritten rules in Her2, Basal (1 rule)
Rewritten rules in LumB, Basal (3 rules)
Rewritten rules in LumA / Her2 (1 / 13 rules each)
Rewritten rules in LumB (2 rules)
Rewritten rules in Basal (11 rules)
ELK1 (TF)

Figure 2: Subtype-specific signaling paths and transcription factors of breast cancer. (a) Overview of the rewritten rules in (a) TNF, (b) TGFB1, (c) EGF pathways. Transcription factors RELA, ATF2 and ELK1 are found in the basal-like-specific path of the pathways, respectively.

subtypes (luminal A, luminal B, HER2-enriched and basal-like) compared to normal cells are determined by `voom` function [47] in Limma R package [39] with a threshold of adjusted P-value ¡ 0.05. Four sets of DEGs were labeled as goals (endpoints) for rule rewriting. The DEGs had few (1 to 3) overlap between subtypes (Supplementary Fig. S1-S3), thereby generating distinctive signaling paths to the DEGs for each subtype (Supplementary Fig. S4-S6 or http://epigenomics.snu.ac.kr/PL/).

3.1 Subtype-specific signaling paths and transcription factors of breast cancer

From TCGA breast cancer data, we found seven different subtype-specific paths originated from three signaling factors (TNF, TGFB1, and EGF) leading to a TF (Table 2). Based on the expression pattern of target genes, we selected four TFs (red texts in Table 2) by testing the subtype-specificity. These TFs (RELA, ATF2, SMAD3, and ELK1) have target genes differentially expressed in the same subtype as their upstream signaling paths. For example, in TNF-RELA case, we found that the ratio of basal-specific DEGs are significantly larger in genes targeted by RELA than non-targeted genes. It shows that RELA regulates downstream target genes in basal-specific manner as well as RELA itself is activated by the basal-specific upstream signal from TNF.

Table 2: Subtype-specific signaling paths and transcription factors of breast cancer detected from PL analysis. Three (ELK1, RELA, ATF2) and one (SMAD3) TFs are detected as subtype-specific regulators for basal-like and luminal B subtypes, respectively.

Pathway	TF	Subtype	# of DEGs	# of total genes	# of DE target genes	# of target genes	Hypergeometric P-value
EGF	**ELK1**	Basal-like	15,676	20,501	1,488	1,886	**0.0038**
TNF	**RELA**	Basal-like	15,676	20,501	387	484	**0.0278**
TGFB1	**ATF2**	Basal-like	15,676	20,501	38	43	**0.0152**
TGFB1	SMAD3	Basal-like	15,676	20,501	477	626	0.5476
TGFB1	**SMAD3**	Luminal B	13,149	20,501	421	626	**0.0445**
TGFB1	SMAD4	Luminal A	13,599	20,501	428	637	0.3071
TGFB1	SMAD4	Luminal B	13,149	20,501	423	637	0.1045

Among the four TFs, RELA, ATF2 and ELK1 are all activated in the basal-like subtype but in the different signaling pathway. Those three paths all started from the common rules of all subtypes but branch off to the basal-specific rules later, mostly through the rules shared with the luminal B subtypes in the midway (Figure 2). Specific rules in each path are shown in Supplementary Figures S7-S10.

3.2 A signaling circuit connecting TNF, TGFB1, and EGF pathways specifically regulated in basal-like breast cancer

From the enrichment test on four TFs (Supplementary Table S1), we identified potential crosstalk partners of EGF, TNF and TGFB1 pathways. For example, in TNF-RELA case, 387 differentially expressed target genes of RELA are significantly enriched with five signaling pathways other than TNF pathway itself, which means that RELA can be considered as a mediator of basal-like specific communication between TNF-pathway and these five external pathways. As a result of the downstream analysis, a potential crosstalk of three pathways (TNF, TGFB1, EGF) by three mediators (RELA, ATF2, ELK1) in the basal-like subtype is discovered and summarized into a signaling circuit (Figure 3). Since a luminal B-specific path of SMAD3 is not involved in the circuit, biological meaning of the path is described in the Supplementary Information.

In the signaling circuit derived for basal-like subtype (Figure 3), we found that three signaling factors (TNF, TGFB1, and EGF) are destined to each other as well as to external pathways. For example, the signal originated from EGF factor comes down to ELK1, then ELK1 broadcasts it to various downstream target genes, one of which is RELA. In turn, RELA propagates the signal to downstream target genes including TGFB1, which subsequently activates ATF2. Finally, ATF2 carries signal to TNF that again activates RELA. This recursively linked signaling cascade constitutes a closed circuit specifically activated in basal-like breast cancers.

Several studies have found that cooperation among the above three pathways induced malignant cancer phenotypes. For example, according to Kali et al., TNF and TGFB1 are found to synergistically increase the cancer stem cell properties in pancreatic cancer cell line [48]. Dunfield et al. also suggest that EGF inhibits the anti-proliferative effect of TGFB1, contributing to increased proliferation in ovarian cancer cells

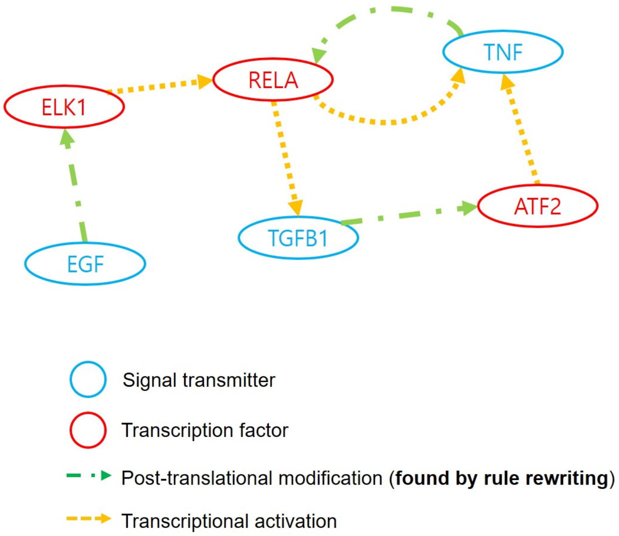
TNF
RELA
ELK1
ATF2
TGFB1
EGF
Signal transmitter
Transcription factor
Post-translational modification (**found by rule rewriting**)
Transcriptional activation

Figure 3: A signaling circuit connecting TNF, TGFB1, and EGF pathways. Three signaling pathways (TNF, TGFB1, and EGF) are connected to each other mediated by three TFs (RELA, ATF2, and ELK1). Blue nodes indicate upstream signaling factor genes, while red nodes indicate TF mediator genes. Green and yellow arrows represents post-translational modification (PTM) and transcriptional regulation, respectively.

[49]. Our system reproduced the close cooperation of three pathways as known in the literature. In addition, our analysis newly discovers three TF mediators as potential links between them.

3.3 Systematical validation of the basal-specific signaling circuit

The signaling circuit of TNF, TGFB1, and EGF pathways is highly associated with basal-like breast cancers, thus we hypothesized that this signaling circuit is a major feature that characterising the basal-like breast cancer. To support the hypothesis, we performed two experiments:

1. A decision tree model was built with the expression values of six genes that belong to the signaling circuit to identify variables characterizing basal-like subtype.
2. For clinical implication, survival analysis was performed using TNF and ELK1 genes that were major discriminatory variables in the decision tree model.

3.3.1 A decision tree model for basal-risk

We built a classification model using six genes in the signaling circuit as input variables and subtype labels (basal/non-basal) of each sample as corresponding target variables. The decision tree (DT) model would show how the combination of six genes contributes to characterising basal-like subtype. To build the model, we used J48 decision tree algorithm implemented in WEKA v3.8 [50] with the confidence factor as 0.25 (default) and the minimum number of objects per leaf as 51, which is approximately 5% of total number of samples (i.e. 1,016 samples).

The primary goal of the DT model is to classify breast cancer samples (1,016 primary tumors in TCGA-BRCA) into several groups having different risk of basal-like subtype. We consider the ratio of samples having basal-like breast cancer in a group of samples as a risk and refer it as *basal-risk*. The result is illustrated in Figure 4. The model was built with gene expression profile normalized with percentile ranked in ascending order among 1,016 samples. The gene expression value of each gene was represented as percentile, where the notation $P_{i\%}$ indicates that the expression level of sample is higher than $i\%$ of 1,016 samples.

We found that the expression level of TNF gene was the most important factor to distinguish basal-like to non-basal-like tumors. Group of samples having TNF expression higher than $P_{90.85\%}$ had 79.57% of

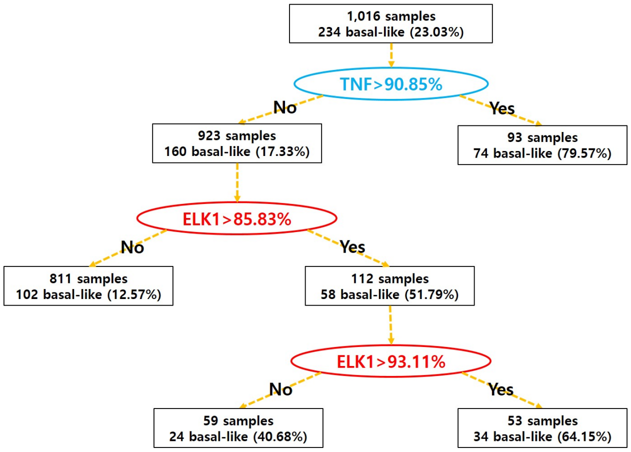

Figure 4: Decision Tree model constructed with six genes in the signaling circuit of basal-like subtype. 1,016 samples from TCGA breast cancer patients are used as gene expression profiles of six genes. Circles indicate decision nodes with the specified condition, classified into the signaling regulators (blue circles) and TFs (red circles). Square boxes show the number of samples and the ratio of basal-like subtype samples in the branch. For example, the top circle written as "TNF¿$P_{90.85\%}$" has two branches among which the right branch with 93 samples satisfy the condition while the left branch includes remaining 923 samples.

basal-risk, which was 4.6 times larger than non-basal-like group. Then, remaining 923 samples was split into two groups by the DT model based on ELK1 expression level, which implied that ELK1 is the second most important factor. In this case, group of samples having ELK1 expression higher than $P_{85.83\%}$ had 51.79% basal-risk, which was 4.1 times larger than non-basal-like group. Lastly, 112 samples having ELK1 expression higher than $P_{85.83\%}$ were split into two groups by the DT model. The group of samples having ELK1 expression higher than $P_{93.11\%}$ had 64.15% of basal-risk, which was 1.6 times larger than non-basal-like group. In summary, the DT model indicates TNF is the most important and dominant factor in deciding whether a sample is basal-like or not. ELK1 was the second important factor and the most dominant factor to samples having TNF expression lower than $P_{90.85\%}$.

3.3.2 Clinical implication of crosstalk between TNF and ELK1 and the contribution of the proposed algorithm

Results from the DT model analysis implies that the patient group with high TNF expression is riskier to be basal-like subtype. When TNF expression is not that high, patient groups with high ELK1 expression still has higher risk to be basal-like (Figure 4). To investigate that the risk of basal-like cancer predictable from the expression of TNF and ELK1 is associated with actual clinical outcome, we performed a survival analysis by drawing Kaplan-Meier plot with measuring log-rank test p-value, using TCGA-BRCA data (Supplementary Figure S11). First, we tested survival difference between TNF-high (¿$P_{90.85\%}$) and TNF-low group (¡=$P_{90.85\%}$), whose log-rank test p-value was 0.034. The result showed that the group of patients having TNF expression higher than $P_{90.85\%}$ had poorer survival outcome, which was consistent with the DT analysis. However, we were not able to find significant survival differences between the three patient groups separated by ELK1 expression at $P_{85.83\%}$ and $P_{93.11\%}$ under TNF $\leq P_{90.85\%}$, whose log-rank test p-value was 0.91. Therefore, we conjecture that TNF expression has a significant association with patient survival as well as basal-risk while the prognostic effect of ELK1 is not conclusive despite its association with basal-risk.

According to Spearman's correlation between expressions of TNF and ELK1, two genes have significant positive correlation (p-value: 4.83E-7). The smoothed plot in Supplementary Figure S11 also shows that TNF expression is strongly associated with ELK1 expression. Moreover, the signaling path deduced by Pathway Logic also indicates that ELK1 is associated with TNF (Figure 3). However, the survival analysis result implies that ELK1 expression level do not have prognostic power. A possible explanation is that ELK1 indirectly increases basal-risk by activating TNF but does not have a direct effect on malignancy by itself. This hypothesis is in part supported by current research findings in that TNF has been considered as a molecular target in breast cancer [3] but no relevant evidence is found for ELK1.

3.4 Performance comparison with existing algorithms

To measure the performance of our proposed system, we analysed the same TCGA breast cancer dataset with two different algorithms: ARACNE and Clipper.

ARACNE [8] was selected as a representative of the reverse engineering algorithm. In the regulatory network predicted by ARACNE, we first filtered the edges with three criteria: 1) both genes in the edge were differentially expressed between cancer subtype and normal samples, 2) the likelihood score of the edge from ARACNE was larger than 0.5 and 3) the type of the edge (activator, suppressor) should be consistent with gene expression information (up- or down-regulated, respectively). Then, we mapped genes in the remaining edges to EGF, TNF, TGFB1 pathways in KEGG [12] database. The results are shown in Supplementary Figures S12-S14. Blue squares denote TFs and red squares denote mapped genes. As shown in the figure, most of the mapped genes were scattered and segmented in the pathways. Some of connected paths in KEGG diagram (e.g. top path in TGFB1, Luminal B subtype) were actually not connected in the ARACNE network. Clipper [14] is an algorithm for detecting signaling paths specific to a specific phenotype. Gene expressions of each cancer subtype and normal samples were analysed by Clipper to identify subtype-specific signal paths. As shown in Supplementary Figures S15-S16, signaling paths detected by Clipper were similar to each other for all subtypes, thus they were not informative for the subtype analysis of cancer.

These experiments showed that signaling pathways are unlikely to be detected from the gene expression difference only, which supports the motivation of our study.

4 Discussion

We proposed a new computational framework to integrate pathway models of PL, gene expression data and transcriptional network to explore subtype-specific biological mechanisms in breast cancer. Models in PL are constructed with detailed descriptions including cellular locations, interactions and PTMs of the molecules. In addition, introducing a transcriptional network into the PL model can lead to the subsequent global changes in a cell such as pathway crosstalk.

Analysing the TCGA breast cancer dataset, we derived paths of signal transduction from key receptors (TNFR, TGFBR1 and EGFR) to key TF regulators (RELA, ATF2, SMAD3 and ELK1) and to downstream genes targeted by TF, specifically detected in a certain subtype. From the basal-specific signaling paths and TFs reported, we derived a pathway crosstalk between three signaling pathways mediated by three TFs, which could provide a novel insight on aggressive breast cancer subtypes. To support the hypothesis that the pathway crosstalk with six genes is the characteristic of basal-like breast cancer, we performed decision tree and survival analysis. The analysis results indicate TNF has strong association with basal-like subtype and bad prognosis, which is supported by the literature, showing the predictive power of the proposed model. ARACNE, a well-known reverse engineering system, is also applied to the same dataset for performance comparison. Regulatory trees of TFs from ARACNE mapped to TNF, TGFB1 and EGF signaling pathways are scattered and shorter than the paths from our methods, showing less efficiency to characterise subtype-specific signaling path.

5 Funding

This research was supported by National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT [No. NRF-2017M3C4A7065887]; The Collaborative Genome Program for Fostering New Post-Genome Industry of the National Research Foundation (NRF) funded by the Ministry of Science and ICT (MSIT) [No. NRF-2014M3C9A3063541]; a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health & Welfare, Republic of Korea [grant number HI15C3224]; and Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) [B0717-16-0098, Development of homomorphic encryption for DNA analysis and biometry authentication].

References

[1] Mahshid Ghoncheh, Zahra Pournamdar, and Hamid Salehiniya. Incidence and mortality and epidemiology of breast cancer in the world. *Asian Pac J Cancer Prev*, 17(Spec No):43–6, 2016.

[2] MF Pang, AM Georgoudaki, L Lambut, J Johansson, V Tabor, K Hagikura, Y Jin, Malin Jansson, JS Alexander, CM Nelson, et al. Tgf-β1-induced emt promotes targeted migration of breast cancer cells through the lymphatic system by the activation of ccr7/ccl21-mediated chemotaxis. *Oncogene*, 35(6):748–760, 2016.

[3] Mingxia Yu, Xiaoxi Zhou, Lin Niu, Guohong Lin, Jin Huang, Wenjing Zhou, Hui Gan, Jing Wang, Xiaodan Jiang, Bingjiao Yin, et al. Targeting transmembrane tnf-α suppresses breast cancer growth. *Cancer research*, 73(13):4061–4074, 2013.

[4] NA Nixon, MB Hannouf, and S Verma. A review of the value of human epidermal growth factor receptor 2 (her2)-targeted therapies in breast cancer. *European Journal of Cancer*, 89:72–81, 2018.

[5] Joel S Parker, Michael Mullins, Maggie CU Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, 27(8):1160, 2009.

[6] A Goldhirsch, William C Wood, Alan S Coates, Richard D Gelber, Beat Thürlimann, H-J Senn, and Panel members. Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. *Ann Oncol*, 22(8):1736–1747, 2011.

[7] Hatice U Osmanbeyoglu, Raphael Pelossof, Jacqueline F Bromberg, and Christina S Leslie. Linking signaling pathways to transcriptional programs in breast cancer. *Genome research*, 24(11):1869–1880, 2014.

[8] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7(1):S7, 2006.

[9] Diana Drago-García, Jesús Espinal-Enríquez, and Enrique Hernández-Lemus. Network analysis of emt and met micro-rna regulation in breast cancer. *Scientific reports*, 7(1):13534, 2017.

[10] Logan A Walsh, Mariano J Alvarez, Erich Y Sabio, Marsha Reyngold, Vladimir Makarov, Suranjit Mukherjee, Ken-Wing Lee, Alexis Desrichard, Şevin Turcan, Martin G Dalin, et al. An integrated systems biology approach identifies trim25 as a key determinant of breast cancer metastasis. *Cell reports*, 20(7):1623–1640, 2017.

[11] Purvesh Khatri, Marina Sirota, and Atul J Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, 8(2):e1002375, 2012.

[12] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[13] Carl F Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow. Pid: the pathway interaction database. *Nucleic acids research*, 37(suppl_1):D674–D679, 2008.

[14] Paolo Martini, Gabriele Sales, M Sofia Massa, Monica Chiogna, and Chiara Romualdi. Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic acids research*, 41(1):e19–e19, 2012.

[15] Ivana Ihnatova, Vlad Popovici, and Eva Budinska. A critical comparison of topology-based pathway analysis methods. *PloS one*, 13(1):e0191154, 2018.

[16] Theresa M Filtz, Walter K Vogel, and Mark Leid. Regulation of transcription factor activity by interconnected post-translational modifications. *Trends in pharmacological sciences*, 35(2):76–85, 2014.

[17] Michelle L Wynn, Nikita Consul, Sofia D Merajver, and Santiago Schnell. Logic-based models in systems biology: a predictive and parameter-free network analysis method. *Integrative biology*, 4(11):1323–1337, 2012.

[18] Luis Mendoza, Denis Thieffry, and Elena R Alvarez-Buylla. Genetic control of flower morphogenesis in arabidopsis thaliana: a logical analysis. *Bioinformatics (Oxford, England)*, 15(7):593–606, 1999.

[19] Luis Mendoza. A network model for the control of the differentiation process in th cells. *Biosystems*, 84(2):101–114, 2006.

[20] Rebekka Schlatter, Kathrin Schmich, Ima Avalos Vizcarra, Peter Scheurich, Thomas Sauter, Christoph Borner, Michael Ederer, Irmgard Merfort, and Oliver Sawodny. On/off and beyond-a boolean model of apoptosis. *PLoS computational biology*, 5(12):e1000595, 2009.

[21] Amit Singh, Juliana M Nascimento, Silke Kowar, Hauke Busch, and Melanie Boerries. Boolean approach to signalling pathway modelling in hgf-induced keratinocyte migration. *Bioinformatics*, 28(18):i495–i501, 2012.

[22] Reza Haydarlou, Annika Jacobsen, Nicola Bonzanni, K Anton Feenstra, Sanne Abeln, and Jaap Heringa. Bioasf: a framework for automatically generating executable pathway models specified in biopax. *Bioinformatics*, 32(12):i60–i69, 2016.

[23] Emek Demir, Michael P Cary, Suzanne Paley, Ken Fukuda, Christian Lemer, Imre Vastrik, Guanming Wu, Peter D'eustachio, Carl Schaefer, Joanne Luciano, et al. The biopax community standard for pathway data sharing. *Nature biotechnology*, 28(9):935, 2010.

[24] G Joshi-Tope, Marc Gillespie, Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Bernard de Bono, Bijay Jassal, GR Gopinath, GR Wu, Lisa Matthews, et al. Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33(suppl_1):D428–D432, 2005.

[25] Ethan G Cerami, Benjamin E Gross, Emek Demir, Igor Rodchenkov, Özgün Babur, Nadia Anwar, Nikolaus Schultz, Gary D Bader, and Chris Sander. Pathway commons, a web resource for biological pathway data. *Nucleic acids research*, 39(suppl_1):D685–D690, 2010.

[26] Steven Eker, Merrill Knapp, Keith Laderoute, Patrick Lincoln, and Carolyn L. Talcott. Pathway Logic: Executable models of biological networks. *Electron Notes Theor Comput Sci*, 71:144–161, 2002.

[27] Manuel Clavel, Francisco Durán, Steven Eker, Patrick Lincoln, Narciso Martí-Oliet, José Meseguer, and Carolyn L. Talcott. *All about Maude - A high-performance logical framework, how to specify, program and verify systems in Rewriting Logic*, volume 4350 of *Lect Notes Comput Sc*. Springer, 2007.

[28] Carolyn Talcott and Merrill Knapp. Explaining response to drugs using pathway logic. In *International Conference on Computational Methods in Systems Biology*, pages 249–264. Springer, 2017.

[29] Johan Rung and Alvis Brazma. Reuse of public genome-wide gene expression data. *Nature Reviews Genetics*, 14(2):89–99, 2013.

[30] Maria Stella Carro, Wei Keat Lim, Mariano Javier Alvarez, Robert J. Bollo, Xudong Zhao, Evan Y. Snyder, Erik P. Sulman, Sandrine L. Anne, Fiona Doetsch, Howard Colman, Anna Lasorella, Ken Aldape, Andrea Califano, and Antonio Iavarone. The transcriptional network for mesenchymal transformation of brain tumours. *Nature*, 463(7279):318–325, Dec 2010.

[31] Juan M Vaquerizas, Sarah K Kummerfeld, Sarah A Teichmann, and Nicholas M Luscombe. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, 10(4):252, 2009.

[32] Uri Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits.* CRC press, 2006.

[33] Cancer Genome Atlas Network and others. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[34] Christina H Stuelten, Stacey DaCosta Byfield, Praveen R Arany, Tatiana S Karpova, William G Stetler-Stevenson, and Anita B Roberts. Breast cancer cells induce stromal fibroblasts to express mmp-9 via secretion of tnf-α and tgf-β. *Journal of cell science*, 118(10):2143–2153, 2005.

[35] Dongwon Baek, Hyun Jin Chun, Dae-Jin Yun, and Min Chul Kim. Cross-talk between phosphate starvation and other environmental stress signaling pathways in plants. *Molecules and cells*, 40(10):697, 2017.

[36] Paul Dent. Crosstalk between erk, akt, and cell survival. *Cancer biology & therapy*, 15(3):245–246, 2014.

[37] Andries Blokzijl, Camilla Dahlqvist, Eva Reissmann, Anna Falk, Annalena Moliner, Urban Lendahl, and Carlos F Ibáñez. Cross-talk between the notch and tgf-β signaling pathways mediated by interaction of the notch intracellular domain with smad3. *The Journal of cell biology*, 163(4):723–728, 2003.

[38] Annette K Larsen, Djamila Ouaret, Karima El Ouadrani, and Amélie Petitprez. Targeting egfr and vegf (r) pathway cross-talk in tumor survival and angiogenesis. *Pharmacology & therapeutics*, 131(1):80–90, 2011.

[39] Gordon K Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

[40] Adrián Riesco, Beatriz Santos-Buitrago, Javier De Las Rivas, Merrill Knapp, Gustavo Santos-García, and Carolyn Talcott. Epidermal growth factor signaling towards proliferation: modeling and logic inference using forward and backward search. *BioMed research international*, 2017, 2017.

[41] Kyuri Jo, Hawk-Bin Kwon, and Sun Kim. Time-series rna-seq analysis package (trap) and its application to the analysis of rice, oryza sativa l. ssp. japonica, upon drought stress. *Methods*, 67(3):364–372, 2014.

[42] Luiz A. Bovolenta, Marcio L. Acencio, and Ney Lemke. HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics*, 13(1):405, Aug 2012.

[43] Jason Ernst, Heather L Plasterer, Itamar Simon, and Ziv Bar-Joseph. Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res*, 20(4):526–536, Apr 2010.

[44] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, Jun 2011.

[45] Xiaohui Xie, Jun Lu, E J Kulbokas, Todd R Golub, Vamsi Mootha, Kerstin Lindblad-Toh, Eric S Lander, and Manolis Kellis. Systematic discovery of regulatory motifs in human promoters and 3'UTRs by comparison of several mammals. *Nature*, 434(7031):338–345, Mar 2005.

[46] Maxim V Kuleshov, Matthew R Jones, Andrew D Rouillard, Nicolas F Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L Jenkins, Kathleen M Jagodnik, Alexander Lachmann, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1):W90–W97, 2016.

[47] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biology*, 15(2):R29, 2014.

[48] Aikyn Kali, Yekaterina O Ostapchuk, and Nikolai N Belyaev. Tnfα and tgfβ-1 synergistically increase the cancer stem cell properties of miapaca-2 cells. *Oncology letters*, 14(4):4647–4658, 2017.

[49] Lesley Deanne Dunfield and Mark William Nachtigal. Inhibition of the antiproliferative effect of tgfβ by egf in primary human ovarian cancer cells. *Oncogene*, 22(30):4745, 2003.

[50] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.