bioRxiv preprint doi: https://doi.org/10.1101/2020.01.25.919639; this version posted July 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Pathway mining in functional genomics: An integrative approach to delineate boolean relationships between Src and its targets

Mehran Piran[1*], Neda Sepahi[1], Mehrdad Piran[1], Pedro L. Fernandes[2], Ali Ghanbariasad[1]

[1] Noncommunicable Diseases Research Center, Fasa University of Medical Sciences, Fasa, Iran

[2] Instituto Gulbenkian de Ciência, Oeiras, Portugal

* piranmehran@gmail.com

Abstract

In recent years the volume of biological data has soared. Parallel to this growth, the need for developing data mining strategies has not met sufficiently. Here we applied data mining techniques on genomic, literature and signaling databases to obtain the required information for pathway inference. An R script was developed in R that discovers pathways using edge information in different signaling databases. We explain how to distinguish more valid pathways from the invalid ones using molecular information in the papers and genomic data analysis. We performed this pathway discovery approach on proto-oncogene c-Src to identify new pathways containing Src and genes that their expression were affected by Src

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.25.919639; this version posted July 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

overactivation. This integrative method uses three sources of information which help bioinformaticians who work with gene regulatory network when they want to infer causal relationships between components of a biological system. In addition, some potential positive and negative feedbacks along with predicted pathways were proposed based on the gene expression results. In fact, this flowchart will open new insights into the interactions between cellular components and help biologists look for new possible molecular relationships that have not been reported neither in signaling databases nor as a pathway.

Author summary:

Since biological systems are an extraordinary complex, the volume of biological data has been soaring exponentially. There are important biological information hidden in the ocean of big data that is achieved by recognizing the true relationships between different sources of data. Since human mind is very limited to find the important molecular relationships, we integrated data mining methods to find new connections between different cellular and molecular components. We illustrated how to utilize genomic data analysis and literature-based study to infer causal relationships between components of a new discovered pathway and distinguish possible true relationships from the false ones. The importance of this method is determined once a researcher intends to develop a large gene regulatory network and this flowchart will help them to firstly, save time and energy secondly, look for any possible relationships. This approach could be useful In complex diseases like cancer and aims to detect new genetic targets to avoid disease enhancement.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.25.919639; this version posted July 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

Data mining tools like programming languages have become an urgent need in the era of big data. Many Biological databases are available for free for the users, but how researchers utilize these repositories depends largely on their computational techniques. Among these repositories, databases in NCBI are of great interest. Pubmed literature database and GEO (Gene Expression Omnibus) (1) and SRA (Sequence Read Archive) (2) genomic databases are among the popular ones.

Different statistical models and machine learning algorithms have been developed to construct a gene regulatory network (GRN) from different genomic and epigenomic data (3, 4). One of the simplest mining approaches to configure a regulatory network is using molecular information in the literature while other researchers use information in genomic repositories for this aim. Many biologists obtain information they need from signaling databases such as KEGG, STRING, OmniPath and so on without considering where this information comes from. In addition, boolean relationships between cellular components suffer from too much simplicity regarding the complex identity of molecular interactions. In our previous work, we demonstrated that there is a week coherency between the gene expression profiles at mRNA level and the sign of interactions coming from signaling databases (5) . A few researchers try to support data they find using different sources of information. Furthermore, many papers illustrate paradox results because of the possible technical biases and type of biological samples. As a result, a deep insight into the type of biological context is needed to infer the correct relationships between cellular components.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.25.919639; this version posted July 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In this study we integrated information coming from literature and genomic data analysis for some sets of constructed pathways. The focus of this study is on the proto-oncogene c-Src which has been implicated in progression and metastatic behavior of human carcinoma and adenocarcinoma (6-9). This gene is engaged in a process called Epithelial to Mesenchymal Transition (EMT) in which epithelial cells lose their cell-cell junction and acquire motility (10). Moreover, it is a target of many anti-cancer drugs, so revealing the new mechanisms that trigger cells to go under EMT would help design more potent drugs for different malignant tumors (11). Two time series transcriptomic datasets were obtained from different technologies, microarray and NGS, in which MCF10A normal human adherent breast cell lines were equipped with ER-Src system. These cells were treated with tamoxifen to witness the overactivation of Src. Gene expression pattern in different time points were analyzed for these cell lines. We tried to find firstly the most affected genes in these cells, secondly all possible connections between Src and DEGs (Differentially Expressed Genes). To be more precise in the analyses, only common DEGs in two datasets were considered. Furthermore, edges information in KEGG and OmniPath databases was used to construct pathways from Src to DEGs and between DEGs themselves. Finally , information in the expression results and the literature were utilized to select the possible correct pathways. Fig 1 illustrates the summary of whole experimental procedure.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.25.919639; this version posted July 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

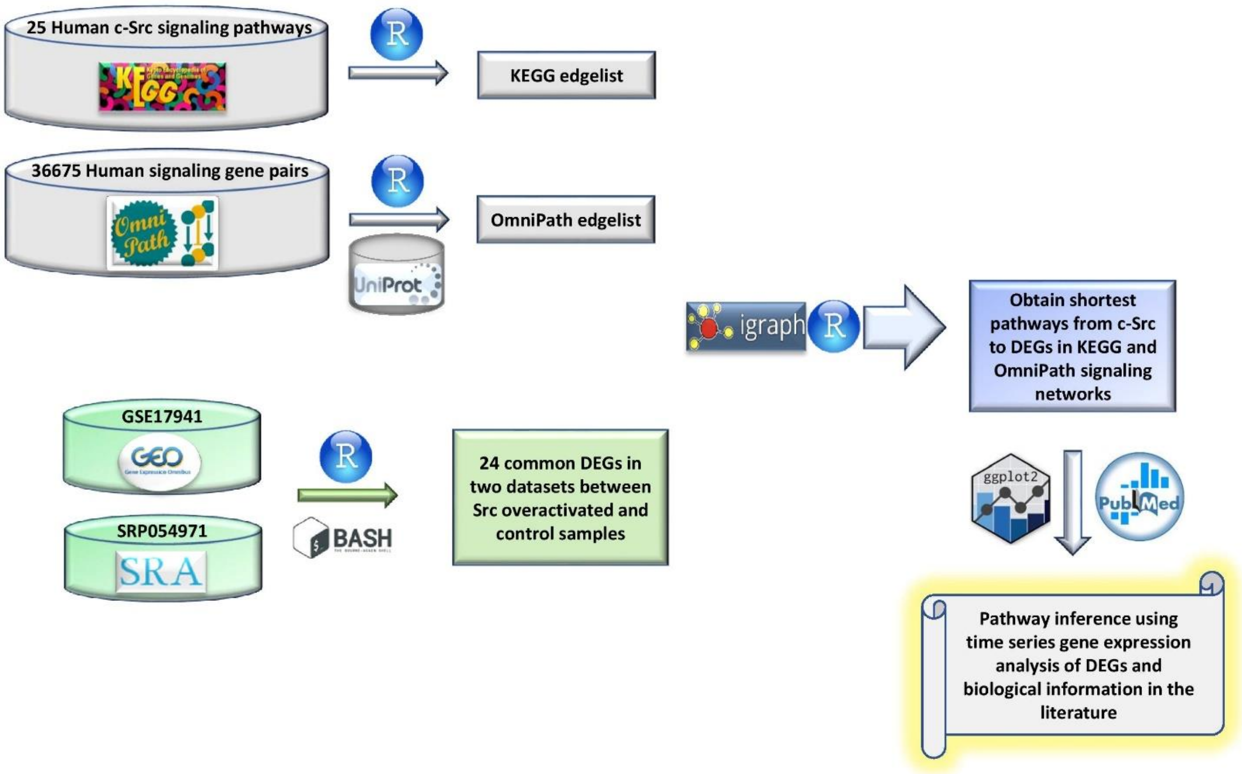

25 Human c-Src signaling pathways
KEGG edgelist
36675 Human signaling gene pairs
OmniPath edgelist
UniProt
igraph
Obtain shortest pathways from c-Src to DEGs in KEGG and OmniPath signaling networks
GSE17941
GEO
SRP054971
SRA
BASH
24 common DEGs in two datasets between Src overactivated and control samples
ggplot2
PubMed
Pathway inference using time series gene expression analysis of DEGs and biological information in the literature

Fig 1: visualization of different datamining steps from genomic and signaling databases to reach signaling pathways. KEGG and OmniPath edgelists were constructed from human KEGG signaling networks containing proto-oncogene c-Src and all human UniProt gene pairs archived in OmniPath database. Two time series gene expression datasets were obtained from GEO and SRA databases and common DEGs were identified in the two datasets between tamoxifen treated samples and ethanol treated samples. Next, a script was developed to recognize all shortest pathways from Src to the DEGs highly affected by Src overactivation. Using time series analysis of DEGs and molecular data about target DEGs in the PubMed, possible valid pathways were distinguished from the false pathways.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.25.919639; this version posted July 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Methods

Database Searching and recognizing pertinent experiments

Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) and Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) databases were searched to detect experiments containing high-quality transcriptomic samples concordance to the study design. Homo sapiens, SRC, overexpression, and overactivation were the keywords used in the search. Microarray raw data with accession number GSE17941 was downloaded from GEO database. RNAseq fastq files with SRP054971 (GSE65885 GEO ID) accession number was downloaded from SRA database. In both studies MCF10A cell lines containing ER-Src system were treated either with tamoxifen to overactivate Src or ethanol as control.

Microarray Data Analysis.

R software version 3.6 was used to import and analyze the data. The preprocessing step involving background correction and probe summarization was done using RMA method in “affy” package (12). Absent probesets were also identified using “mas5calls” function in this package. If a probeset contained more than two absent values, that one was regarded as absent and removed from the expression matrix. Besides, outlier samples were identified and removed using PCA and hierarchical clustering approaches. Employing Quantile normalization method, data were normalized. Many to Many problem which is mapping multiple probesets to the same gene symbol, was resolved using nsFilter function in genefilter package (13). This function selects the probeset with the largest interquartile range (IQR) to be representative of other probesets mapping to the same gene symbol. After

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.25.919639; this version posted July 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

that, “limma” R package was utilized to identify differentially expressed genes between tamoxifen treated and ethanol treated cells (14).

RNA-seq Data Analysis

Seven samples of RNAseq study useful for the aim of our research were selected and their SRA IDs were from SRX876039 to SRX876045 (15). Bash shell commands were used to reach from fastq files to the counted expression files. Quality control step was done on each sample separately using “fastqc” function in “FastQC” module (16). Next, “Trimmomatic” software was used to trim reads (17). 10 bases from the reads head were cut and bases with quality less than 30 were removed and reads with the length of larger than 36 base pair were kept. Then, trimmed files were aligned to the hg38 standard FASTA reference sequence using “HISAT2” software to create SAM files (18). SAM files converted into BAM files using “Samtools” package (19). In the last step, BAM files and a GTF file containing human genome annotation were given to “featureCounts” program to create counted files (20). After that, files were imported into R software (v3.6) and all samples were attached together to construct an expression matrix with 59,412 variables and seven samples. Rows with sum values less than 7 were removed from the expression matrix and RPKM method in edger R package was used to normalize the rest of the rows (21). In order to identify DEGs, “limma” R package was used (14). Finally, correlation-based hierarchical clustering was done using “factoextra” R package [14].

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.25.919639; this version posted July 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Pathway Construction from Signaling Databases

25 human KEGG signaling networks containing Src element were downloaded from KEGG signaling database (22). Pathways were imported into R using "KEGGgraph" package (23) and using programming techniques, all the pathways were combined together. Loops were omitted and only directed inhibition and activation edges were selected so a large KEGG edgelist was constructed. In addition, a very large edgelist containing all literature curated mammalian signaling pathways were constructed from OmniPath database (24). To do pathway discovery, a script was developed in R using igraph package (25) by which a function was created with four arguments. The first argument accepted an edgelist, second argument was a vector of source genes, third argument was a vector of target genes and forth argument received a maximum length of pathways.

Results

Data Preprocessing and Identifying Differentially Expressed Genes

Almost 75% of the probesets were regarded as absent and left out from the expression matrix in order to avoid any technical errors. To be more precise in the preprocessing step, outlier sample detection was conducted using PCA (using eigenvector 1 (PC1) and eigenvector 2 (PC2)) and hierarchical clustering. Fig 2A illustrates the PCA plot for the samples in GSE17941 study. Sample GSM448818 in time point 36-hour, was far away from the other samples which might be an outlier sample. In the hierarchical clustering approach, Pearson correlation coefficients between samples were subtracted from one for measurement of the distances. Then, samples were plotted based on their Number-SD. To

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.25.919639; this version posted July 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

get this number for each sample, the average of whole distances was subtracted from average of distances in all samples, then results of these subtractions were divided by the standard deviation of distance averages (26). Sample GSM448818_36h with Number-SD less than negative two was regarded as the outlier and removed from the dataset (Fig 2B).

There were 21 upregulated and 3 downregulated common DEGs between the two datasets. Fig 3 illustrates the average expression values between two groups of tamoxifen-treated samples and ethanol-treated samples for these common DEGs. For the RNAseq dataset (SRP054971) average of 4-hours, 12-hour and 24-hour time points were used (A) and for the microarray dataset (GSE17941) average of 12-hour and 24-hour time points were utilized (B). All DEGs had the absolute log fold change larger than 0.5 and p-value less than 0.05. Housekeeping genes are situated on the diagonal of the plot whilst all DEGs are located above or under the diagonal. This demonstrates that the preprocessed datasets were of sufficient quality for the analysis. In both datasets SERPINB3 was the most upregulated gene.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.25.919639; this version posted July 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

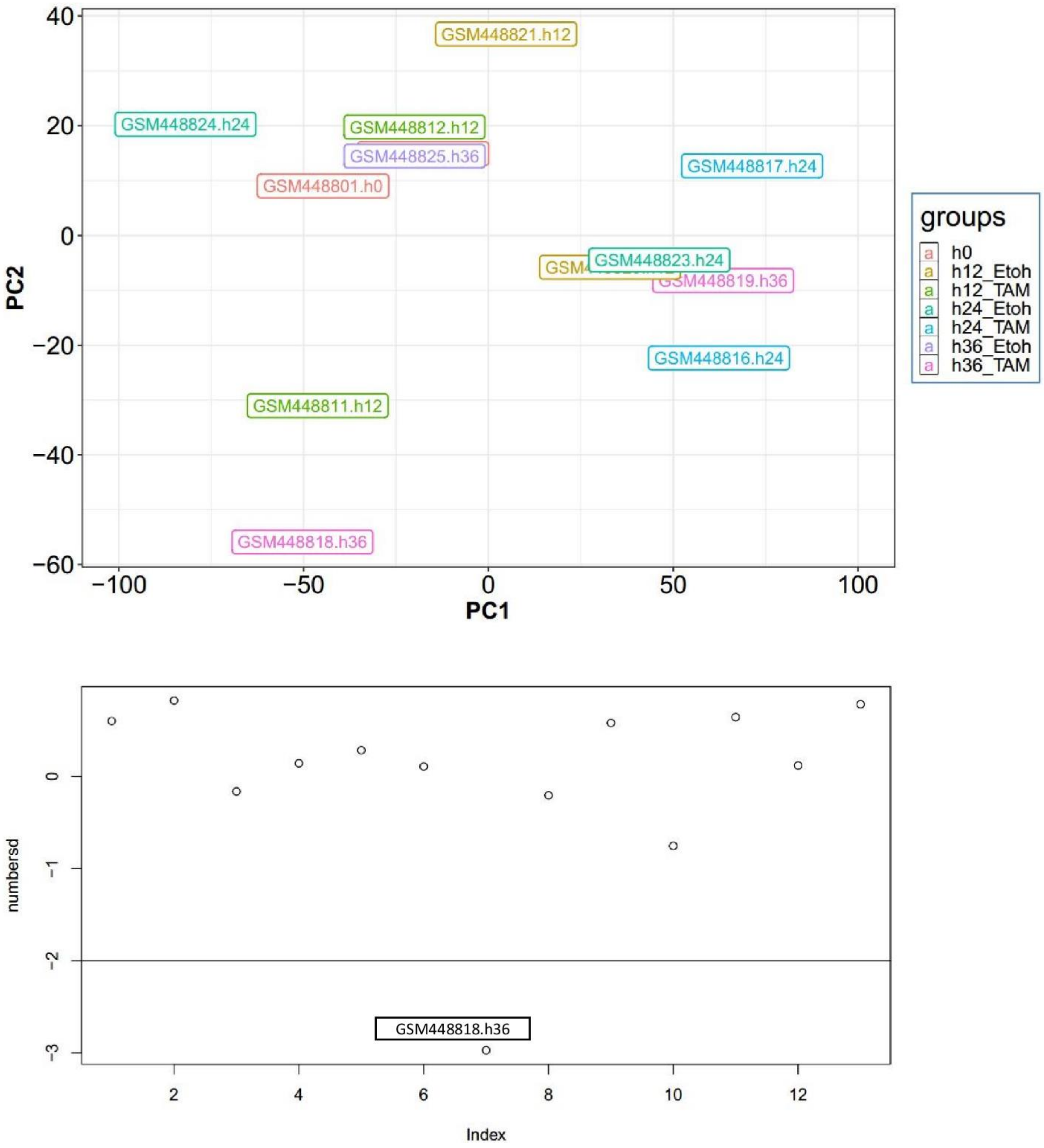


Fig 2: Outlier detection: A, is the PCA between samples in defferent time points in GSE17941 dataset. Replicates are in the same color. B, illustrates the numbersd value for each sample. Samples under -2 are regarded as outlier. The x axis represents the indices of samples.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.25.919639; this version posted July 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

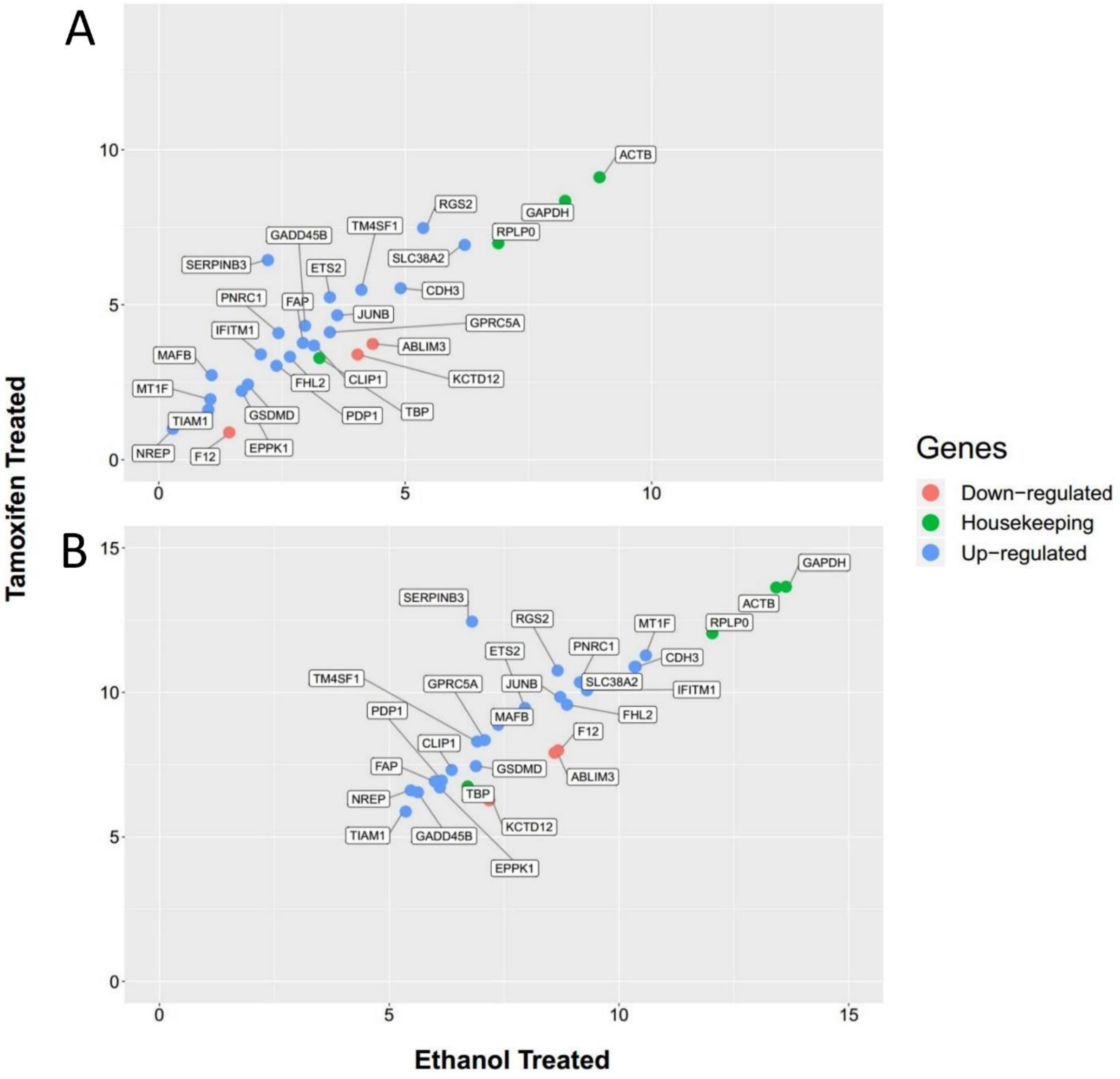


Fig 3: Scatter plot for upregulated, downregulated and housekeeping genes. The average values in different time points in SRP054971, A, and GSE17941, B, datasets were plotted between control (ethanol treated) and tamoxifen treated.

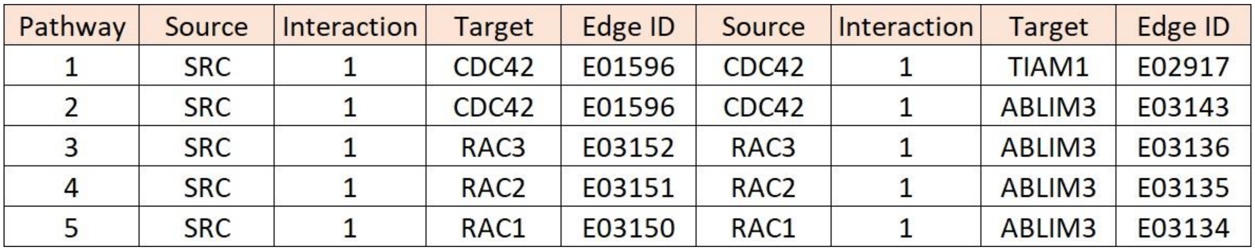
Pathway Analysis

Because obtained common DEGs were the results of analyzing datasets from two different genomic technologies, there was a high probability that Src-activation influence expression

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.25.919639; this version posted July 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of these 24 genes. As a result, we continued our analysis in KEGG signaling network to find the Boolean relationships between Src and these genes. To this end, we developed a pathway mining approach which extracts all possible directed pathways between two components in a signaling network using "igraph" R programming tool. An R script was written to construct a large edgelist containing 1025 nodes and 7008 edges from 25 downloaded pathways containing Src (supplementary file 1). To reduce the number of pathways, only shortest pathways were regarded in the analysis. Just two DEGs were present in the constructed KEGG network namely TIAM1 and ABLIM3. Table1 shows all the pathways between Src and these two genes. They were two-edge distance Src targets in which their expression was affected by Src overactivation. TIAM1 was upregulated while ABLIM3 was down-regulated in Fig 3. In Pathway number 1, Src induces CDC42 and CDC42 induces TIAM1 respectively. Therefore, a total positive interaction is yielded from Src to TIAM1. On the one hand ABLIM3 was down-regulated in Fig 3, but on other hand this gene could positively be induced by Src activation in four different ways in Table 1. We proposed that these pathways are not valid at mRNA level. Moreover, information in the signaling databases comes from different experimental sources and biological contexts. Consequently, a precise biological interpretation is required when combining information from different studies to predict a regulatory pathway.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.25.919639; this version posted July 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

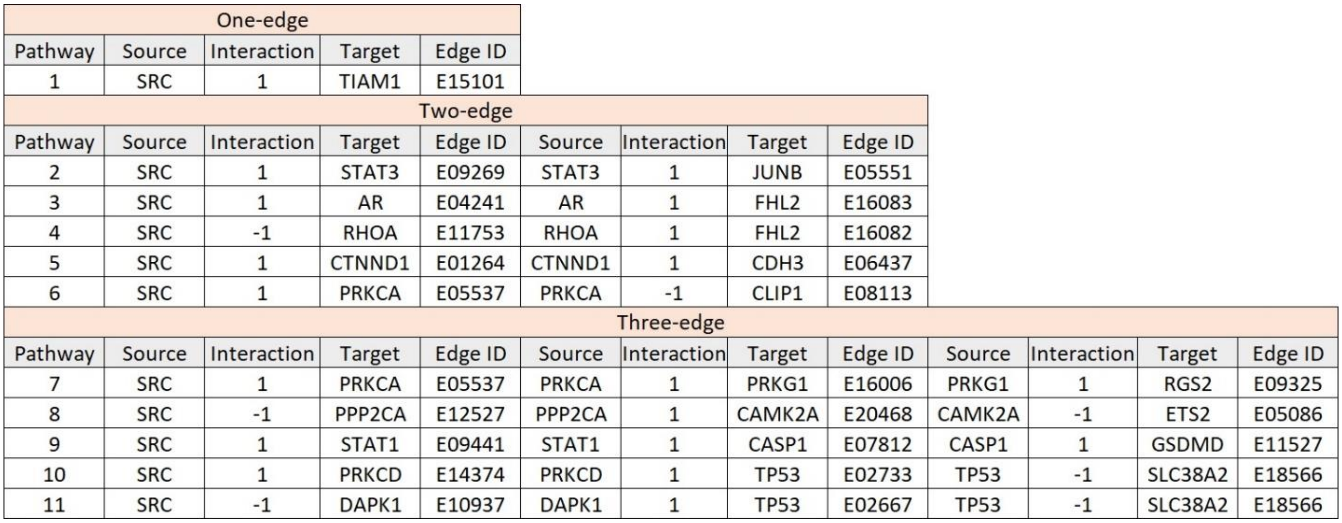


Pathway	Source	Interaction	Target	Edge ID	Source	Interaction	Target	Edge ID
1	SRC	1	CDC42	E01596	CDC42	1	TIAM1	E02917
2	SRC	1	CDC42	E01596	CDC42	1	ABLIM3	E03143
3	SRC	1	RAC3	E03152	RAC3	1	ABLIM3	E03136
4	SRC	1	RAC2	E03151	RAC2	1	ABLIM3	E03135
5	SRC	1	RAC1	E03150	RAC1	1	ABLIM3	E03134

Table 1: Discovered Pathways from Src gene to TIAM1 and ABLIM3 genes. In the interaction column, "1" presents activation state and "-1" shows inhibition state. ID column presents the edge IDs (indexes) in the KEGG edgelist. Pathway column represents the pathway number.

Due to the uncertainty about the discovered pathways in KEGG, a huge human signaling edgelist was constructed from OmniPath database (http://omnipathdb.org/interactions) utilizing R programming. Constructed edgelist was composed of 20853 edges and 4783 nodes. Fourteen DEGs were found in the edgelist eleven of them were recognized to be Src targets. Eleven pathways with the maximum length of three and minimum length of one were discovered illustrated in Table2. Unfortunately, ABLIM1 gene was not found in the edgelist but Tiam1 was a direct (one-edge distance) target of Src in Pathway 1. FHL2 could be induced or suppressed by Src based on Pathways 3 and 4 respectively. However, FHL2 was among the upregulated genes by Src. Therefore, the necessity of biological interpretation of each edge is required to discover a possible pathway.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.25.919639; this version posted July 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



One-edge												
Pathway	Source	Interaction	Target	Edge ID								
1	SRC	1	TIAM1	E15101								
Two-edge												
Pathway	Source	Interaction	Target	Edge ID	Source	Interaction	Target	Edge ID				
2	SRC	1	STAT3	E09269	STAT3	1	JUNB	E05551				
3	SRC	1	AR	E04241	AR	1	FHL2	E16083				
4	SRC	-1	RHOA	E11753	RHOA	1	FHL2	E16082				
5	SRC	1	CTNND1	E01264	CTNND1	1	CDH3	E06437				
6	SRC	1	PRKCA	E05537	PRKCA	-1	CLIP1	E08113				
Three-edge												
Pathway	Source	Interaction	Target	Edge ID	Source	Interaction	Target	Edge ID	Source	Interaction	Target	Edge ID
7	SRC	1	PRKCA	E05537	PRKCA	1	PRKG1	E16006	PRKG1	1	RGS2	E09325
8	SRC	-1	PPP2CA	E12527	PPP2CA	1	CAMK2A	E20468	CAMK2A	-1	ETS2	E05086
9	SRC	1	STAT1	E09441	STAT1	1	CASP1	E07812	CASP1	1	GSDMD	E11527
10	SRC	1	PRKCD	E14374	PRKCD	1	TP53	E02733	TP53	-1	SLC38A2	E18566
11	SRC	-1	DAPK1	E10937	DAPK1	1	TP53	E02667	TP53	-1	SLC38A2	E18566

Table 2: Discovered pathways from Src to the DEGs in OmniPath edgelist. In the interaction column, "1" depicts activation state and "-1" exhibits inhibition state. ID column presents the edge IDs (indexes) in the OmniPath edgelist. Pathway column represents the pathway number.

Time Series Gene Expression Analysis

Src was overactivated in MCF10A (normal breast cancer cell line) cells using Tamoxifen treatment at the time points 0-hour, 1-hour, 4-hour and 24-hour in the RNAseq dataset and 0-hour, 12-hour, 24-hour and 36-hour in the microarray study. The expression values for all upregulated genes in Src-activated samples were higher than controls in all time points in both datasets. The expression values for all downregulated genes in Src-activated samples were less than controls in all time points in both datasets. Fig 4 depicts the expression values for TIAM1, ABLIM3, RGS2, and SERPINB3 at these time points. RGS2 and SERPINB3 witnessed a significant expression growth in tamoxifen-treated cells at the two datasets. The time-course expression patterns of all DEGs are presented in supplementary file 2.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.25.919639; this version posted July 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

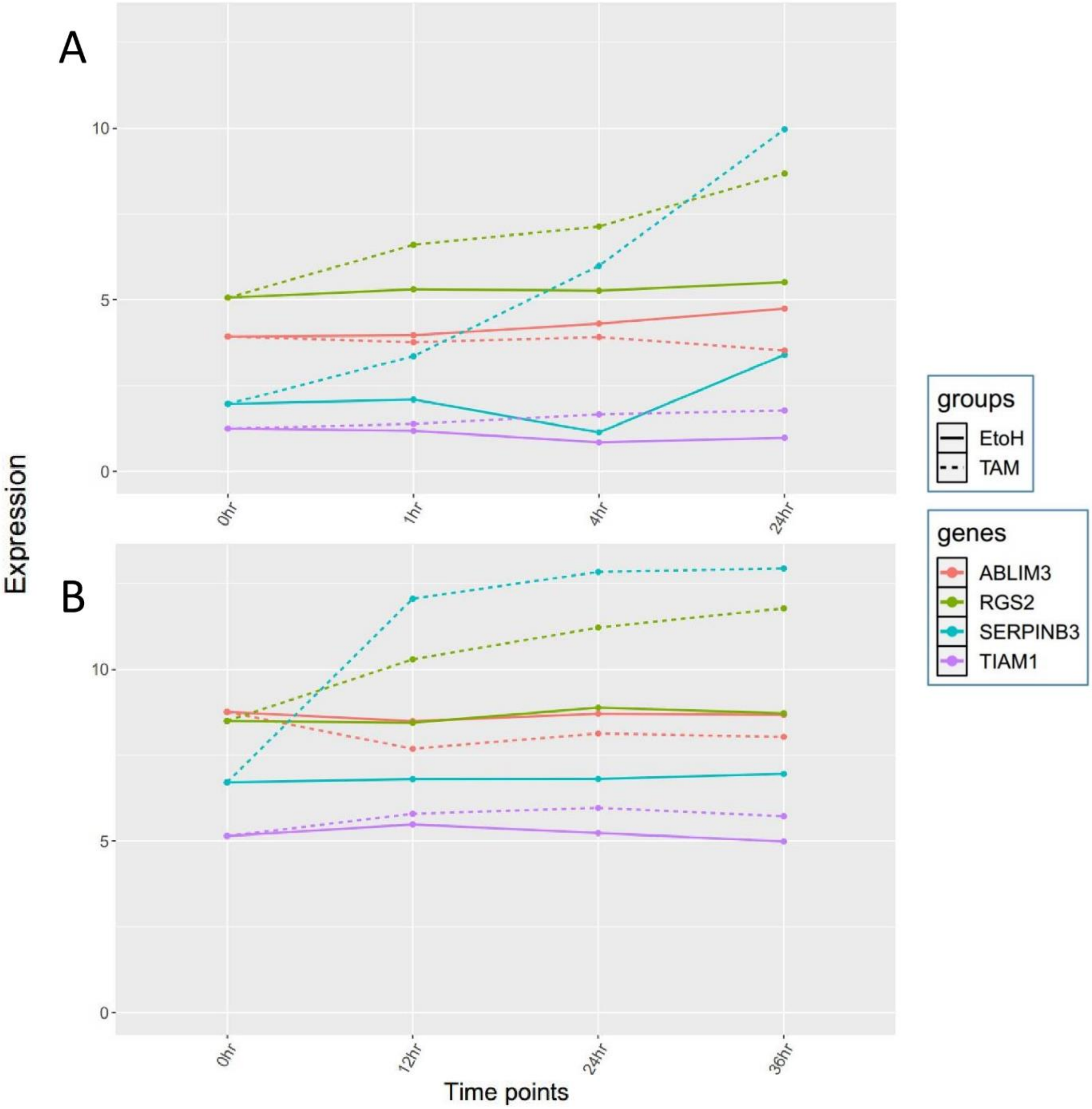


Fig 4: Expression values related to the four DEGs. TIAM1 and ABLIM3 are the genes present in the KEGG signaling network and RGS2 and SERPINB3 are the genes highly affected by activation of Src. A, presents values in SRP054971 dataset. B, presents values in GSE17941 dataset.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.25.919639; this version posted July 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Clustering

We applied hierarchical clustering on expression of all DEGs just in Src over-activated samples. Pearson correlation coefficient was used as the distance in the clustering method. Clustering results were different between the two datasets, therefore, we applied this method only on SRP054971 Dataset (Fig 5). We hypothesized that genes in close distances within each cluster may have relationships with each other. Among the four DEGs in Fig 4, Only TIAM1 and RGS2 were present in OmniPath edgelist. As a result, pathways were extracted from TIAM1 and RGS2 to their cluster counterparts. All these pathways are presented in supplementary file 3.

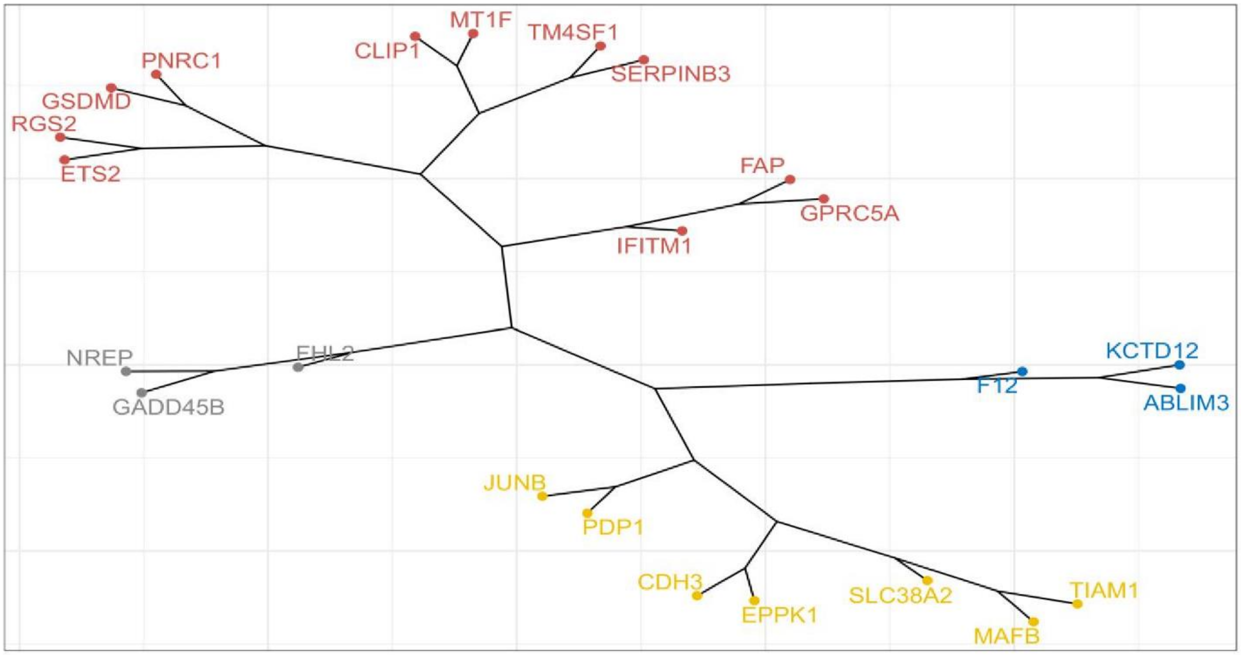


Fig 5: Pearson correlation-based hierarchical clustering. Figure shows the clustering results for expression of DEGs in RNA-seq dataset. Four clusters were emerged members of each are in the same color.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.25.919639; this version posted July 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Discussion

In many pathway inference methods, the necessity of valid data is required. Our approach is simple, is not based on mathematical models but real evidence. This pathway inference flowchart would demonstrate the importance of post transcriptional data and might also help distinguish the components in a pathway affected at transcript level and those affected at post-transcriptional level. Therefore, if other types of data rather than only mRNA is considered, many new biological pathways would be predicted by connecting edge information in more signaling databases. For instance, proteomic and phospho-proteomic data will be needed along with a Protein-Protein Interaction Network (PPIN) if transition of the signals are related to the phosphorylation process (27-29). Methylation sites on both mRNA and protein impact cell signaling (30). Other modifications such as acetylation (31-33), ubiquitylation (34), sumoylation (35) impact the activity of different signaling networks as well. As a result, different layers of gene regulation need to be implemented on each gene pair once inferring causal relationships between the gene pairs.

Analyzing the relationships between Src and all of these 24 obtained DEGs were of too much biological information. Therefore, we concentrated on TIAM1 and ABLIM3 (Actin binding LIM protein family member 3) and two highly affected genes in tamoxifen-treated cells namely RGS2 and SERPINB3 presented in Fig 4. More investigations are needed to be done on the rest of the genes to see how and why all these genes are affected by Src. That's important because numerous studies have reported the role of Src implication in promoting EMT (36-39).

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.25.919639; this version posted July 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

TIAM is a guanine nucleotide exchange factor (GEF) that is phosphorylated in tyrosine residues in cells transfected with oncogenic Src (40). Therefore, it could be a direct target of Src which has not been reported in 25 KEGG networks but this relationship was found in Table 2, pathway number 1. Src induces cell transformation through both Ras-ERK and Rho family of GTPases dependent pathways (41, 42). Moreover, The Rho GTPase Cdc42 can be phosphorylated and activated by Src through epidermal growth factor (EGF) signaling (43). CDC42 itself can promote activation of c-Src through EGFR signaling (44). Therefore, a positive feedback might occurs between the two molecules by Src activation. Not only CDC42 but also RAC1,2,3 GTPases are activated by Src through oncogenic growth factor receptors. These activations are transduced to GEF proteins such as TIAM1 which in turn regulate Rho-like GTPases GDP/GTP exchange so keep GTPases in their active form (45, 46). Therefore, by following the track of information in pathway number one in Tables 1 and 2, firstly a consistency in the flow of information is emerged secondly, the validity of pathways were demonstrated. Furthermore, Expression induction of TIAM1 in Fig 4 would explain how c-Src bolsters its cooperation with TIAM1 in order to keep Rho family of GTPases active leading to formation of membrane ruffles in vivo (40).

ABLIM3 is a component of adherent junctions (AJ) that interacts with actin filaments in epithelial cells and hepatocytes. Its downregulation in Fig 4 leads to the weakening of cell-cell junctions (47). Information flow for pathways targeting ABLIM3 is contrary to the results in Fig 4, therefore literature should be investigated carefully and probably post-transcriptional data is required to be analyzed, so we can predict that pathways numbers 2 to 5 in Table 1 are invalid during EMT in MCF10A cell lines.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.25.919639; this version posted July 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SERPINB3 is a serine protease inhibitor that is overexpressed in epithelial tumors to inhibit apoptosis. This gene induces deregulation of cellular junctions by suppression of E-cadherin and enhancement of cytosolic B-catenin (48). In the Hypoxic environment of hepatic tumors, SERPINB3 is upregulated by HIF-2α and this upregulation requires intracellular generation of ROS (49). Moreover, there is a positive feedback between SERPINB3 and HIF-1α and -2α in liver cancer cells (50). Therefore, these positive mechanisms would help cancer cells to augment invasiveness properties and proliferation. Unfortunately, this gene did not exist neither in OmniPath nor in KEGG edgelists. Its significant expression induction and its effects in promoting metastasis would explain one of the mechanisms that Src triggers invasive behaviors in cancer cells.

RGS2 is a GTPase activating protein (GAP) for the alpha subunit of heterotrimeric G proteins (51). Although this gene was highly upregulated in Src-overexpressed cells, its expression has been found to be reduced in different cancers such as prostate and colorectal cancer (52, 53). This might overcomplicate the oncogenic effects of Src on promoting cancer. Pathway number 7 connects Src to RGS2 in Table 2. To infer pathway 7, ErbB2-mediated cancer cell invasion in breast cancer happens through direct interaction and activation of PRKCA/PKCα by Src (54). PKG1/PRKG1 is phosphorylated and activated by PKCα following phorbol 12-myristate 13-acetate (PMA) treatment (55). PRKG1 is a serine-threonine kinase activated through GMP binding. Inhibition of phosphoinositide (PI) hydrolysis in smooth muscles is done via phosphorylation of RGS2 by PRKG and its association with Gα-GTP subunit of G proteins. In fact, PRKG over-activates RGS2 to accelerate Gα-GTPase activity and enhance Gαβγ (complete G protein) trimer formation (56). Consequently, there would be a possibility that Src can induce RGS2 protein activity regarding the edge information in

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.25.919639; this version posted July 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

pathway 7. Src overactivation significantly induced expression of RGS2 in Fig 4. Therefore, there should be another pathway that increase the expression of RGS2 or maybe this upregulation is mediated by the mentioned components. Furthermore, RGS2 activates the GTPase activity of G proteins, so the upregulation of RGS2 by Src help to activate G proteins and promotion of cellular transformation.

SERPINB3 and ABLIM3 were not present in the OmniPath edgelist, so we conducted pathway mining from TIAM1 and RGS2 to their cluster counterparts and between TIAM1 and RGS2 (supplementary file 3). The results show that there would be a possible relationship between PNRC1, ETS2 and FATP toward RGS2. All these relationships were set by PRKCA and PRKG genes at the end of pathways. Nevertheless, PNCR1 made shorter pathways and is worth more investigation. JUNB and PDP1 make a relationship with TIAM1. The discovered pathway from JUNB to TIAM1 was mediated by EGFR and SRC demonstrating that SRC could induce its expression by induction of JUNB. Moreover, there were two pathways from TIAM1 to RGS2 and from RGS2 to TIAM1 with the same length which are worth more investigation.

In summary, Pathway mining is not just looking for direct pathways. Many hidden relationships will be unfolded once all findings and results in different databases around a specific component come together to reach a target component. Likewise we discovered possible new arrange of relationships between Src and Tiam1, between Src and RGS2 and between DEGs themselves.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.25.919639; this version posted July 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research. 2002;30(1):207-10.
2. Leinonen R, Sugawara H, Shumway M, Collaboration INSD. The sequence read archive. Nucleic acids research. 2010;39(suppl_1):D19-D21.
3. Mochida K, Koda S, Inoue K, Nishii R. Statistical and machine learning approaches to predict gene regulatory networks from transcriptome datasets. Frontiers in Plant Science. 2018;9:1770.
4. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, et al., editors. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC bioinformatics; 2006: Springer.
5. Piran M, Karbalaei R, Piran M, Aldahdooh J, Mirzaie M, Ansari-Pour N, et al. Can we assume the gene expression profile as a proxy for signaling network activity? Biomolecules. 2020;10(6):850.
6. Finn R. Targeting Src in breast cancer. Annals of Oncology. 2008;19(8):1379-86.
7. Gargalionis AN, Karamouzis MV, Papavassiliou AG. The molecular rationale of Src inhibition in colorectal carcinomas. International journal of cancer. 2014;134(9):2019-29.
8. Irby RB, Yeatman TJ. Role of Src expression and activation in human cancer. Oncogene. 2000;19(49):5636.
9. Kim LC, Song L, Haura EB. Src kinases as therapeutic targets for cancer. Nature reviews Clinical oncology. 2009;6(10):587.
10. Yang J, Weinberg RA. Epithelial-mesenchymal transition: at the crossroads of development and tumor metastasis. Developmental cell. 2008;14(6):818-29.
11. Shukla D, Meng Y, Roux B, Pande VS. Activation pathway of Src kinase reveals intermediate states as targets for drug design. Nature communications. 2014;5(1):1-11.
12. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3):307-15.
13. Gentleman R, Carey V, Huber W, Hahne F, Maintainer MBP, AnnotationDbi I. Package 'genefilter'. 2013.
14. Smyth GK, Ritchie M, Thorne N, Wettenhall J, Shi W, Hu Y. limma: Linear Models for Microarray and RNA-Seq Data User's Guide. 2002.
15. Ji Z, Song R, Regev A, Struhl K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. elife. 2015;4:e08890.
16. Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.
17. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.
18. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature protocols. 2016;11(9):1650.
19. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
20. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923-30.
21. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-40.
22. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000;28(1):27-30.
23. Zhang JD, Wiemann S. KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. Bioinformatics. 2009;25(11):1470-1.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.25.919639; this version posted July 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

24. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nature methods. 2016;13(12):966.
25. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal, Complex Systems. 2006;1695(5):1-9.
26. Oldham MC, Horvath S, Konopka G, Iwamoto K, Langfelder P, Kato T, et al. Identification and Removal of Outlier Samples Supplement for:" Functional Organization of the Transcriptome in Human Brain. dim (dat1).1(18631):105.
27. Tang J, Gautam P, Gupta A, He L, Timonen S, Akimov Y, et al. Network pharmacology modeling identifies synergistic Aurora B and ZAK interaction in triple-negative breast cancer. npj Systems Biology and Applications. 2019;5(1):20.
28. D'Souza RC, Knittle AM, Nagaraj N, van Dinther M, Choudhary C, Ten Dijke P, et al. Time-resolved dissection of early phosphoproteome and ensuing proteome changes in response to TGF-β. Sci Signal. 2014;7(335):rs5-rs.
29. Hill SM, Heiser LM, Cokelaer T, Unger M, Nesser NK, Carlin DE, et al. Inferring causal molecular networks: empirical assessment through a community-based effort. Nature methods. 2016;13(4):310.
30. Biggar KK, Li SS-C. Non-histone protein methylation as a regulator of cellular signalling and function. Nature reviews Molecular cell biology. 2015;16(1):5-17.
31. Mittal R, Peak-Chew S-Y, McMahon HT. Acetylation of MEK2 and IκB kinase (IKK) activation loop residues by YopJ inhibits signaling. Proceedings of the National Academy of Sciences. 2006;103(49):18574-9.
32. Lin Z, Guo H, Cao Y, Zohrabian S, Zhou P, Ma Q, et al. Acetylation of VGLL4 regulates Hippo-YAP signaling and postnatal cardiac growth. Developmental cell. 2016;39(4):466-79.
33. Song Y, Li Z-x, Liu X, Wang R, Li L-w, Zhang Q. The Wnt/β-catenin and PI3K/Akt signaling pathways promote EMT in gastric cancer by epigenetic regulation via H3 lysine 27 acetylation. Tumor Biology. 2017;39(7):1010428317712617.
34. Wertz IE, Dixit VM. Signaling to NF-κB: regulation by ubiquitination. Cold Spring Harbor perspectives in biology. 2010;2(3):a003350.
35. Lee PS, Chang C, Liu D, Derynck R. Sumoylation of Smad4, the common Smad mediator of transforming growth factor-β family signaling. Journal of Biological Chemistry. 2003;278(30):27853-63.
36. Nagathihalli NS, Merchant NB. Src-mediated regulation of E-cadherin and EMT in pancreatic cancer. Front Biosci (Landmark Ed). 2012;17:2059-69.
37. Wilson C, Nicholes K, Bustos D, Lin E, Song Q, Stephan J-P, et al. Overcoming EMT-associated resistance to anti-cancer drugs via Src/FAK pathway inhibition. Oncotarget. 2014;5(17):7328.
38. Zhao Y, Li X, Sun X, Zhang Y, Ren H. EMT phenotype is induced by increased Src kinase activity via Src-mediated caspase-8 phosphorylation. Cellular physiology and biochemistry: international journal of experimental cellular physiology, biochemistry, and pharmacology. 2012;29(3-4):341-52.
39. Zhou J, Li X, Wang F, Ren M, Du M. Effects of c-Src kinase on lens diseases associated with EMT of human lens epithelial cells. 2019.
40. Bustelo XR. Rac1 function is required for Src-induced transformation: Evidence of a role for Tiam1 and Vav2 in Rac activation by Src. 2003.
41. Leng R, Liao G, Wang H, Kuang J, Tang L. Rac1 expression in epithelial ovarian cancer: effect on cell EMT and clinical outcome. Medical oncology. 2015;32(2):28.
42. Timpson P, Jones GE, Frame MC, Brunton VG. Coordination of cell polarization and migration by the Rho family GTPases requires Src tyrosine kinase activity. Current biology. 2001;11(23):1836-46.
43. Shen Y, Hirsch DS, Sasiela CA, Wu WJ. Cdc42 regulates E-cadherin ubiquitination and degradation through an epidermal growth factor receptor to Src-mediated pathway. Journal of Biological Chemistry. 2008;283(8):5127-37.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.25.919639; this version posted July 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

44. Tu S, Wu WJ, Wang J, Cerione RA. Epidermal growth factor-dependent regulation of Cdc42 is mediated by the Src tyrosine kinase. Journal of Biological Chemistry. 2003;278(49):49293-300.
45. del Mar Maldonado M, Dharmawardhane S. Targeting rac and Cdc42 GTPases in cancer. Cancer research. 2018;78(12):3101-11.
46. Liu L, Zhao L, Zhang Y, Zhang Q, Ding Y. Proteomic analysis of Tiam1-mediated metastasis in colorectal cancer. Cell biology international. 2007;31(8):805-14.
47. Matsuda M, Yamashita JK, Tsukita S, Furuse M. abLIM3 is a novel component of adherens junctions with actin-binding activity. European journal of cell biology. 2010;89(11):807-16.
48. Quarta S, Vidalino L, Turato C, Ruvoletto M, Calabrese F, Valente M, et al. SERPINB3 induces epithelial–mesenchymal transition. The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland. 2010;221(3):343-56.
49. Cannito S, Turato C, Paternostro C, Biasiolo A, Colombatto S, Cambieri I, et al. Hypoxia up-regulates SERPINB3 through HIF-2α in human liver cancer cells. Oncotarget. 2015;6(4):2206.
50. Cannito S, Villano G, Turato C, Morello E, Foglia B, Novo E, et al. SerpinB3 up-regulates hypoxia inducible factors-1α and-2α in liver cancer cells through different mechanisms. Digestive and Liver Disease. 2016;48:e19.
51. Cunningham ML, Waldo GL, Hollinger S, Hepler JR, Harden TK. Protein kinase C phosphorylates RGS2 and modulates its capacity for negative regulation of Gα11 signaling. Journal of Biological Chemistry. 2001;276(8):5438-44.
52. Jiang Z, Wang Z, Xu Y, Wang B, Huang W, Cai S. Analysis of RGS2 expression and prognostic significance in stage II and III colorectal cancer. Bioscience reports. 2010;30(6):383-90.
53. Linder A, Thulin MH, Damber J-E, Welén K. Analysis of regulator of G-protein signalling 2 (RGS2) expression and function during prostate cancer progression. Scientific reports. 2018;8(1):17259.
54. Tan M, Li P, Sun M, Yin G, Yu D. Upregulation and activation of PKC α by ErbB2 through Src promotes breast cancer cell invasion that can be blocked by combined treatment with PKC α and Src inhibitors. Oncogene. 2006;25(23):3286-95.
55. Hou Y, Lascola J, Dulin NO, Richard DY, Browning DD. Activation of cGMP-dependent protein kinase by protein kinase C. Journal of Biological Chemistry. 2003;278(19):16706-12.
56. Nalli AD, Kumar DP, Al-Shboul O, Mahavadi S, Kuemmerle JF, Grider JR, et al. Regulation of Gβγ i-dependent PLC-β3 activity in smooth muscle: inhibitory phosphorylation of PLC-β3 by PKA and PKG and stimulatory phosphorylation of Gα i-GTPase-activating protein RGS2 by PKG. Cell biochemistry and biophysics. 2014;70(2):867-80.

Supporting information:

S1: KEGG networks and KEGG and OmniPath edgelists

S2: Time-series gene expression data

S3: discovered pathways in OmniPath signaling database