Pathway Mining and Data Mining in Functional Genomics. An Integrative Approach to Delineate Boolean Relationships between Src and Its Targets

Mehran Piran[1,2*], Pedro L. Fernandes[2], Neda Sepahi[3,4], Mehrdad Piran[5,6], Amir Rahimi[1]

[1] Bioinformatics and Computational Biology Research Center, Shiraz University of Medical Sciences, Shiraz, Iran

[2] Instituto Gulbenkian de Ciência, Oeiras, Portugal

[3] Department of Medical Biotechnology, Fasa University of Medical Sciences, Fasa, Iran

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

[5] Department of Tissue Engineering and Applied Cell Sciences, School of Advanced Technologies in Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran

[6] Department of Biology, East Tehran Branch, Islamic Azad University, Tehran, Iran

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. Bioinformaticians utilize different data mining techniques to obtain the required information they need from genomic, transcriptomic and proteomic databases to construct a gene regulatory network (GNR). One of the simplest mining approaches to construct a GNR is reading a great number of papers to configure a large network (for instance, a network with 50 nodes and 200 edges) which takes a lot of time and energy. Here we introduce an integrative method that combines information from transcriptomic data with sets of constructed pathways. A program was written in R that makes pathways from edgelists in different signaling databases. Furthermore, we explain how to distinguish false pathways from the correct ones using literature study or incorporating the biological information with gene expression results. This approach can help bioinformaticians and mathematical biologists who work with GRNs when they want to infer causal relationships between components of a biological system. Once you know which direction to go, you will reduce the number of false results.

Introduction

Data mining tools like programming languages has become an urgent need in the era of big data. The more important issue is how and when these tools are required to be implemented. Many Biological databases are available free for the users, but how researchers utilize these repositories depends on their computational techniques. Among these repositories, databases in NCBI are of great interest. GEO (Gene Expression Omnibus) [1] and SRA (Sequence Read Archive) [2] are two important genomic databases that archive microarray and next generation sequencing (NGS) data. Pubmed (US National Library of Medicine) is the library database in NCBI that stores most of the published papers in the area of biology and medicine. Biologists and computational biologists utilize these databases depend on their aim of study. Many researchers use information in the literature to construct a GNR. While other researchers use information in genomic repositories. Some of them obtain the information they want from signaling databases such as KEGG, STRING, OmniPath and so on without considering where this information comes from. A few researchers try to support data they find using other sources of information. Furthermore, many papers are present in the literature which illustrates paradox results. They contain molecular techniques data such as Real-time PCR, Immunoprecipitation, western blot and high-throughput techniques. As a result, a deep insight into the type of biological context is needed to choose the correct answer.

In this study we propose an integrative method to use information in genomic repositories, signaling databases and literature using the knowledge of data mining and computational techniques and knowledge of molecular biology. This study was concentrated on proto-oncogene c-Src which has been implicated in progression and metastatic behavior of human cancers including those of colon and breast [3-6]. Moreover, this gene is a target of many chemotherapy drugs, so revealing the new mechanisms that trigger cells to go under Epithelial to Mesenchymal Transition (EMT) would help design more pertinent drugs for different carcinomas and adenocarcinomas.

Boolean relationships between molecular components of cells suffer from too much simplicity regarding the complex identity of molecular interactions. In the meanwhile, many interactions can be connected together to offer a pathway. To reach from an oncogene to a

bioRxiv preprint first posted online Jan. 25, 2020; doi: http://dx.doi.org/10.1101/2020.01.25.919939. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

junctional protein, millions of pathways can be configured if only information for most reported genes is considered. To determine the right ones among thousands of pathways, they should be validated based on the type of biological context and valid experimental results. Therefore, in this study, two transcriptomic datasets were utilized 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 and gene expression pattern in different time points were analyzed in these cell lines. Because during EMT process many junctional proteins or related proteins and kinases are deregulated in expression or activation, we tried to find the most affected genes in these cells and all possible connections between Src and DEGs. To be more precise about the result of analysis, we found two transcriptomic datasets obtained from two different technologies, microarray and NGS. Then we only considered DEGs which were common in both datasets. Moreover, we used information in KEGG and OmniPath databases to construct pathways from Src to DEGs (Differentially Expressed Genes) and between DEGs themselves. Then information from expression results and those from the papers were utilized to select the possible correct pathways.

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, overactivation are the keywords utilized in the search. Microarray raw data with accession number GSE17941 was downloaded from GEO database. RNA-seq raw data with SRP054971 (GSE65885 GEO ID) accession number was downloaded from SRA database. In both studies, either tamoxifen was used to overactivate Src or ethanol was used as control in in MCF10A cell line containing ER-Src system.

bioRxiv preprint first posted online Jan 25, 2020; doi: http://dx.doi.org/10.1101/2020.01.25.916838. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Microarray Data Analysis.

R software was used to import and analyze the data for each dataset separately. The preprocessing step involving background correction and probe summarization was done using RMA method in Affy package [7]. 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. Next, data were normalized using quantile normalization method. Many to Many problem which is mapping multiple probesets to the same gene symbol was solved using nsFilter function in genefilter package [8]. This function selects the probeset with the highest interquartile range (IQR) as the representative of other probesets mapping to the same gene symbol. After that, LIMMA R package was utilized to identify differentially expressed genes (DEGs) [9].

RNA-seq Data Analysis

Samples are related to the study where they illustrated that many pseudogenes and lncRNAs are translated into functional proteins [10]. We selected seven samples of this study useful for the aim of our research and their SRA IDs are from SRX876039 to SRX876045. Bash shell commands in Ubuntu Operating System were used to reach counted expression files from fastq files. Quality control step was done on each sample separately using fastqc function in FastQC module. Next, Trimmomaticsoftware was used to trim reads. 10 bases from the reads head were cut and bases with quality less than 30 were removed from the reads and reads with the length of larger than 36 base pair (BP) were kept. Then, trimmed files were aligned to the hg38 standard fasta reference sequence using HISAT2 software to create SAM files. SAM files converted into BAM files using Samtools package. In the last step, BAM files and a GTF file containing human genome annotation were given to featureCounts program to create counted files. After that, files were imported into R software and all samples were attached together and an expression matrix constructed with 59412 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 remaining rows [11]. Finally,in

bioRxiv preprint first posted online Jan. 25, 2020; doi: http://dx.doi.org/10.1101/2020.01.21.919633. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

order to identify DEGs, LIMMA R package was used. Finally, correlation-based hierarchical clustering was done using factoextra r package [14].

Pathway Construction from Signaling Databases

25 human KEGG signaling networks containing Src element were downloaded from KEGG signaling database [12]. Pathways were imported into R using "KEGGgraph" package [13] and using programming techniques, all the pathways were combined together. Loops were omitted and only directed inhibition and activation edges were selected. In addition, a very large edgelist containing all literature curated mammalian signaling pathways were constructed from OmniPath database [14]. To do pathway discovery, a script was developed in R using igraph package [15] by which a function was created with four arguments. The first argument accepts an edgelist, second argument is a vector of source genes, third argument is a vector of target genes and forth argument receives a maximum length of pathways.

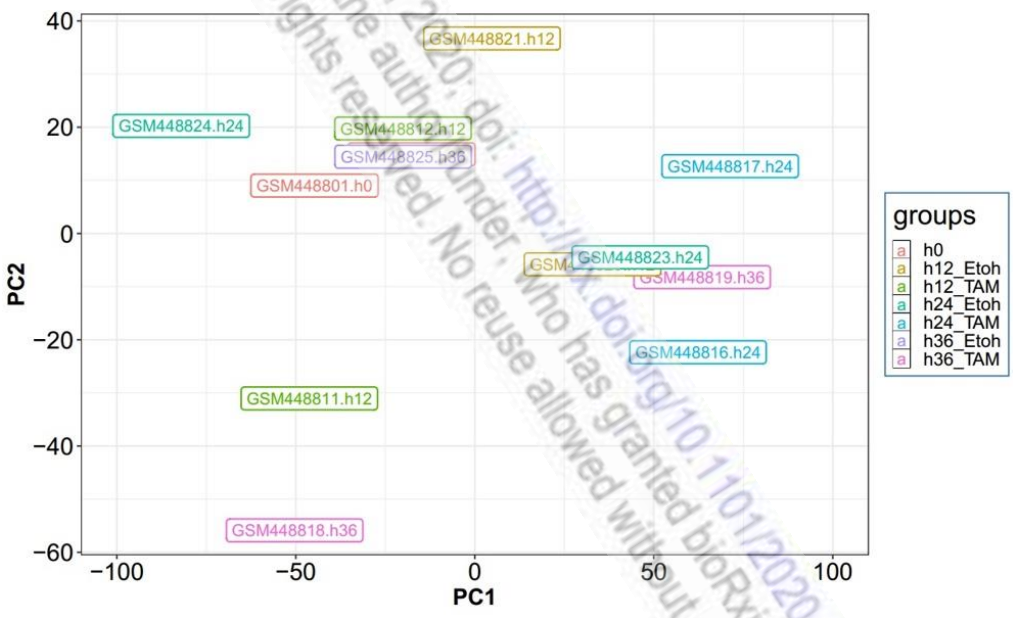
Results

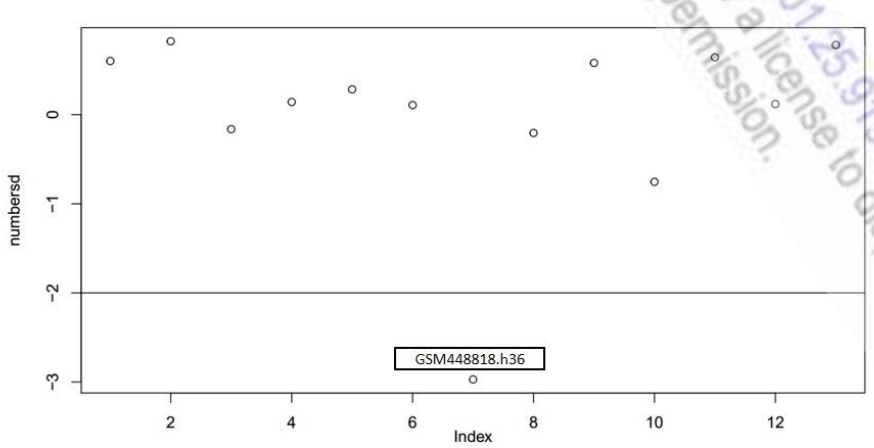
Data Preprocessing and Identifying Differentially Expressed Genes

Almost 75% of probesets were regarded as absent and left out from the expression matrix to avoid 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. Figure1A illustrates the PCA plot for the samples in GSE17941 study. Sample GSM448818 in time point 36-hour, was far away from the other samples. 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 get this number for each sample, the average of whole distances is subtracted from distances average in all samples, then results of these subtractions are normalized (divided) by the standard deviation of distance averages [16]. Sample GSM448818_36h with Number-SD less than negative two was regarded as the outlier and removed from the dataset (Figure 1B). There were 21 upregulated and 3 downregulated

bioRxiv preprint first posted online Jan 25, 2021; doi: http://dx.doi.org/10.1101/2020.01.25.919969 The copyright holder for this preprint (which was not peer reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

common DEGs between two datasets. Figure 2 illustrates the average expression values between two groups of tamoxifen-treated samples and ethanol-treated samples for these common DEGs at different time points. 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 have 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 are of sufficient quality for the analysis.





bioRxiv Preprint first posted online Jan. 25, 2020; doi: http://dx.doi.org/10.1101/2020.01.25.919639. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figure1: Outlier detection, A is the PCA between samples in defferent time points in GSE17941 dataset. Replicates are in the same color. B illustrate the numbersd value for each sample. Samples under -2 are regarded as outlier. The x axis represents the indices of samples.

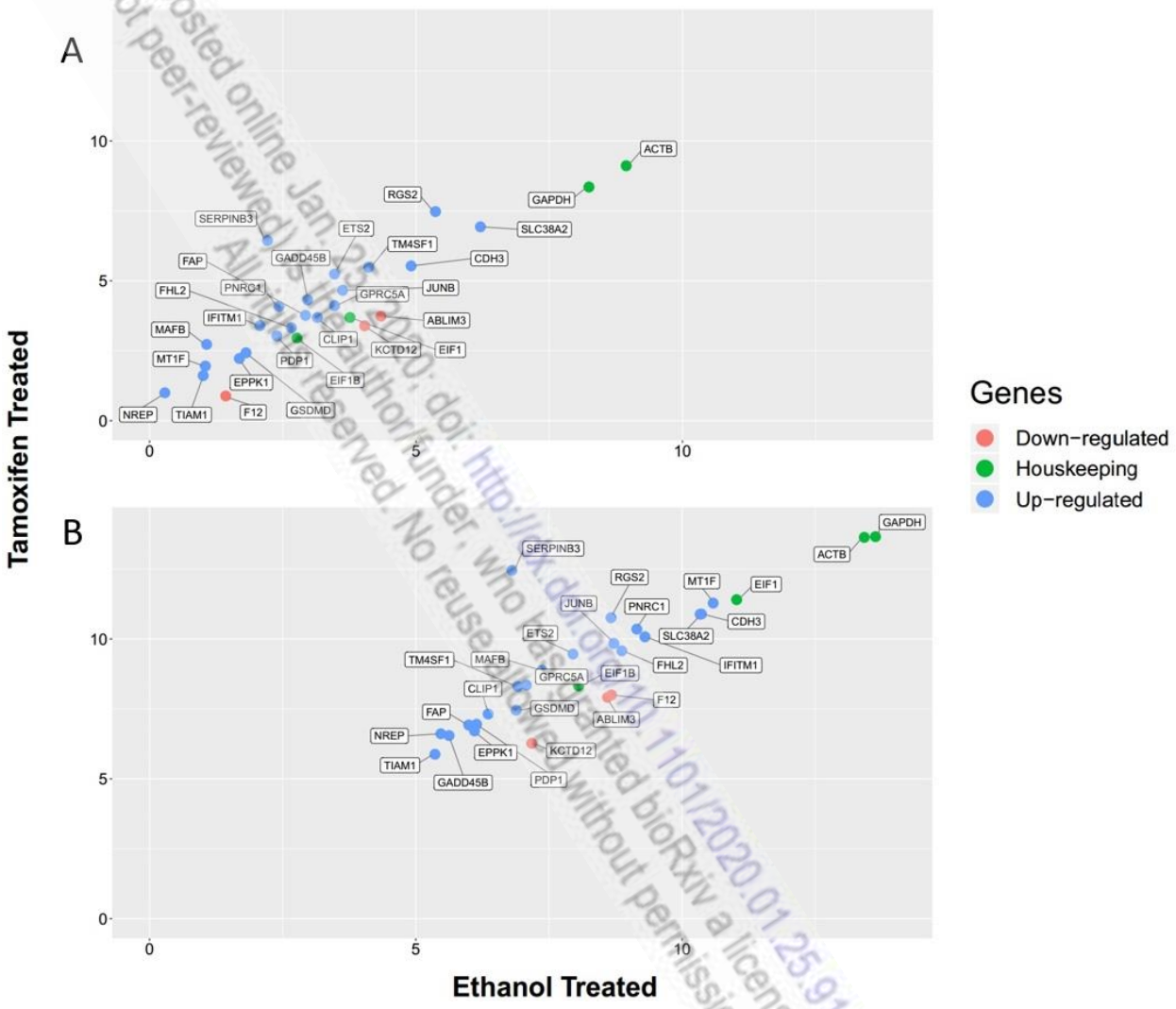


Figure 2: Scatter plot for upregulated, downregulated and housekeeping genes. The average values in different time points in SRP054971, A, and GSE17941, B, datasets were plotted between control (ethanol treated) and tamoxifen treated.

Pathway Mining

Since, the obtained DEGs are the results of analyzing datasets from two different genomic technologies, there is a high probability of affecting Src activation on these 24 genes. So, we

bioRxiv preprint first posted online Jan. 26, 2020; doi: http://dx.doi.org/10.1101/2020.01.25.919696. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

continued our analysis in KEGG signaling networks containing Src to find the Boolean relationships between Src and these genes. To this end, we developed a pathway mining approach which extracts all possible pathways between two components in a signaling network. For pathway discovery, a large edgelist with 1025 nodes and 7008 edges was constructed from 25 downloaded pathways (supplementary file 1). To reduce number of pathways, only shortest pathways were regarded in the analysis. Just two DEGs were presented in the constructed KEGG network namely TIAM1 and ABLIM3. Table1 shows all the pathways between Src and these two genes. They are two-edge distance Src targets which their expression is affected by Src over-activation. TIAM1 was upregulated while ABLIM3 was down-regulated in Figure 2. In Pathway number 1, Src induces CDC42 and CDC42 induces TIAM1. Therefore, a total positive interaction is yielded from Src to TIAM1. On the one hand ABLIM3 was down-regulated in Figure 2, On other hand this gene could positively be induced by Src activation in four different ways in Table 1. This paradox results, firstly would be a demonstration that Boolean interpretation of relationships suffers from lack of reality which simplifies the system artificially and ignores all the kinetics of interactions. Secondly, 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 construct a gene regulatory network.

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

Table1: Discovered Pathways from Src gene to TIAM1 and ABLIM3 genes. In the interaction column, "1" means activation and "-1" means inhibition. 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).

bioRxiv preprint first posted online Jan 25, 2020; doi: https://doi.org/10.1101/2020.01.25.919639. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Constructed edgelist is composed of 20853 edges and 4783 nodes. 14 DEGs were found in the edgelist which eleven of them were found to be Src targets. 11 Shortest pathways with maximum length of three and minimum length of one were discovered illustrated in Table2. Unfortunately, ABLIM1 gene was not found in the edgelist. So, we couldn't further its pathway analysis with Src. But Tiam1 would be 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 is among the upregulated genes by Src. Therefore, the necessity of biological interpretation of each edge is required to discover the correct pathway.

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

Table2: Discovered pathways from Src to the DEGs in OmniPath edgelist.

Time Series Gene Expression Analysis

Src was over-activated 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 value for all upregulated genes in Src-activated samples was higher than controls in all time points in both datasets. The expression value for all downregulated genes in Src-activated samples were less than controls in all time points in both datasets. Figure 3 depicts the expression values for TIAM1, ABLIM3, RGS2, and SERPINB3 in these time points. RGS2 and SERPINB3

bioRxiv preprint first posted online Jan. 25, 2020 at https://doi.org/10.1101/2020.01.25.919639. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

witnessed a significant expression growth in tamoxifen-treated cells at the two datasets. Time-course expression patterns of all DEGs are presented in supplementary file 2.

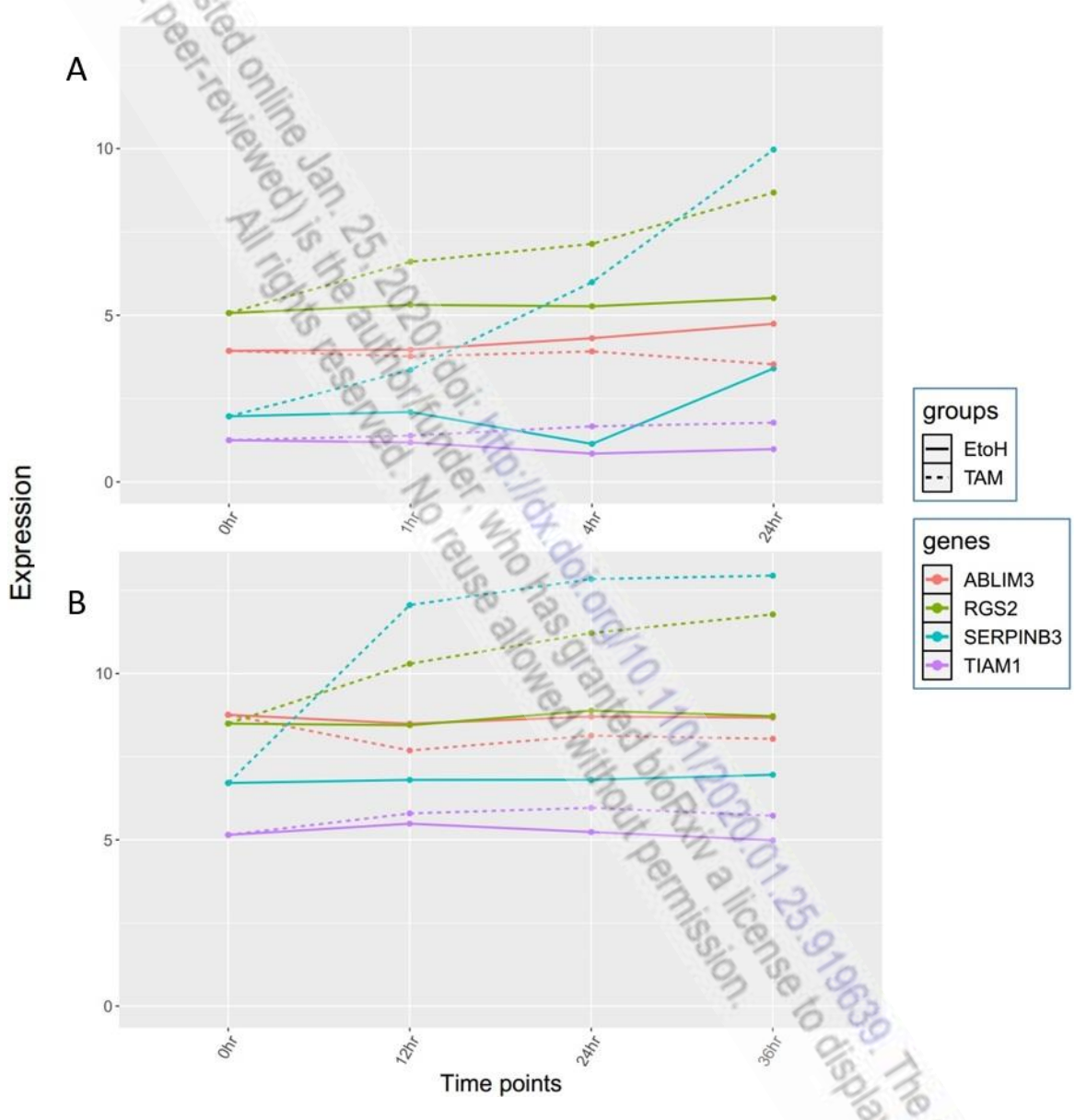


Figure 3: Expression values related to four DEGs.

bioRxiv preprint first posted online Jan. 25, 2020; doi: http://dx.doi.org/10.1101/2020.01.25.919639. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. 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 (Figure4). We hypothesized that genes in close distances in each cluster may have relationships with each other, so we applied pathway analysis on four DEGs in Figure 3. Among them, Only TIAM1 and RGS2 were present in OmniPath edgelist. As a result, we extracted pathways from TIAM1 and RGS2 to their cluster counterparts. All these pathways are presented in supplementary file 3.

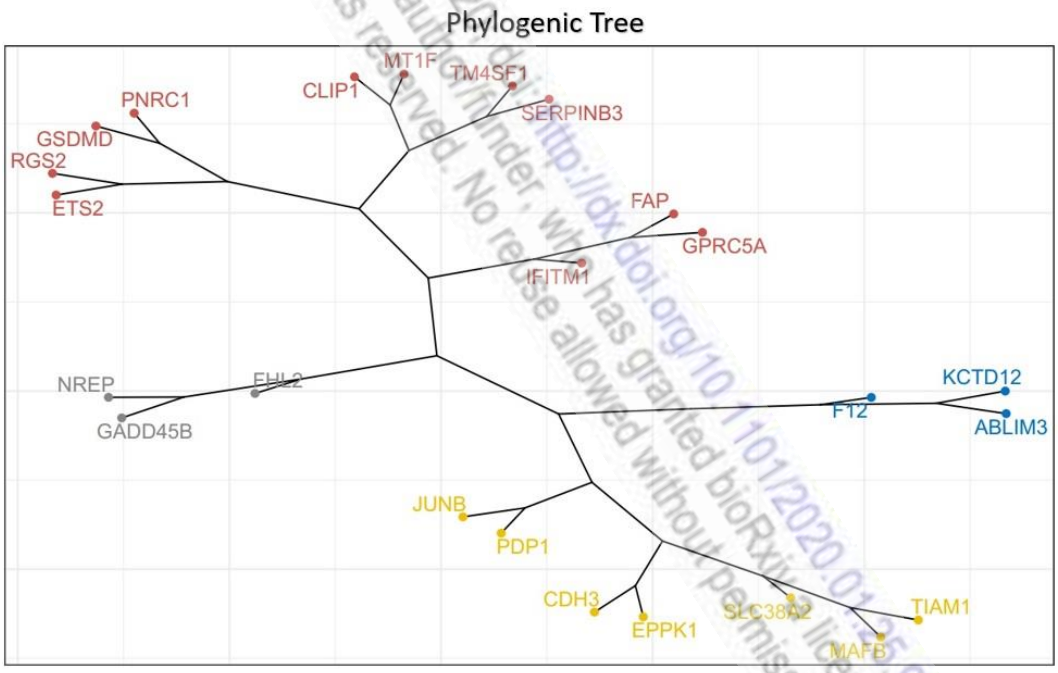


Figure 4: Correlation-based hierarchical clustering. Figure shows the clustering results for expression of DEGs in RNA-seq dataset. Four clusters were emerged members of each have the same color.

Discussion

Analyzing the relationships between Src and all these 24 obtained DEGs were of too much biological information and also were beyond the goal of this study, Therefore, we considered only TIMA1 and

bioRxiv preprint first posted online Jan. 25, 2020; doi: http://doi.org/10.1101/2020.01.25.919039. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

ABLIM3 and two highly affected genes in tamoxifen-treated cells namely RGS2 and SERPINB3 presented in Figure3. So, 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 of some reports in multiple studies that Src alone is not sufficient to totally promote EMT []. Importantly, Src is the target of many chemotherapy drugs, so evaluate the function of affected genes is vital for drug design.

ABLIM3 (Actin binding LIM protein 1) is a component of adherent junctions (AJ) in epithelial cells and hepatocytes. Its down-regulation in Figure3 leads to the weakening of cell-cell junctions [17]. Rac1 is a mall GTP-binding protein (G protein) that its activity is strongly elevated in Src-transformed cells. It is a critical component of the pathways connecting oncogenic Src with cell transformation [18] which also was presented by edge E03150 in pathway number 5 in table1. Tiam1 is a guanine nucleotide exchange factor that is phosphorylated in tyrosine residues in cells transfected with oncogenic Src. Therefore, it could be a direct target of Src which has not been reported in 25 KEGG networks but this relationship was found in Table2 pathway number 1. Tiam1 cooperates with Src to induce activation of Rac1 in vivo and formation of membrane ruffles. Expression induction of TIAM1 in Figure 3 would explain how Src bolster its cooperation with TIAM1 to induce EMT and cell migration [19]. Src induces cell transformation through both Ras-ERK and Rho family of GTPases dependent pathways [20, 21]. Rac1 downregulation decreases cell EMT and proliferation capability in vitro and in vivo [20].

Par polarity complex (Par3–Par6–aPKC) regulates the establishment of cell polarity. CDC42 and Rac control the activation of the Par polarity complex. TIAM1 is an activator of Rac and a crucial component of the Par complex in regulating epithelial (apical–basal) polarity [22]. Based on the KEGG human Chemokine signaling pathway with map ID hsa04062, this complex is activated by CDC42. So, regarding the information in pathway number 7, there is a possibility that Src-induced activation of Tiam1 is also mediated by CDC42.

SERPINB3 is a serine protease inhibitor that is over-expressed in epithelial tumors to inhibit apoptosis. This gene induces deregulation of cellular junctions by suppression of E-cadherin and enhancement of cytosolic B-catenin supported by features of epithelial–mesenchymal transition [23]. Hypoxia up-regulates SERPINB3 through HIF-2α in human liver cancer cells and this up-regulation under hypoxic conditions requires intracellular generation of ROS [24]. Moreover, there is a positive feedback that is SERPINB3 up-regulates HIF-1α and -2α in liver cancer cells [25]. 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.

bioRxiv preprint first posted online Jan. 25, 2023; doi: https://doi.org/10.1101/2023.01.25.919839. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Its significant expression induction and its mentioned mechanism in promoting metastasis would explain one of the mechanisms that Src triggers invasive behaviors in cancer cells.

Regulator of G protein 2 or in short RGS2 is a GTPase activating protein (GAP) for G alpha subunits of heterotrimeric G proteins. Increasing the GTPase activity of G protein alpha subunit drives their inactive GDP binding form [26]. Although this gene was highly up-regulated in Src-overexpressed cells, its expression has been found to be reduced in different cancers such as prostate and colorectal cancer [27, 28]. This might be one of the contradictory effects of Src on promoting EMT. In Table 2, Pathway number 7 connects Src to RGS2. ErbB2-mediated cancer cell invasion in breast cancer samples is practiced through direct interaction and activation of PRKCA/PKCα by Src [29]. PKG1/PRKG1 is phosphorylated and activated by PKC following phorbol 12-myristate 13-acetate (PMA) treatment [30]. GMP binding activates PRKG1, which phosphorylates serines and threonines on many cellular proteins. 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 [31]. Consequently, there would be a possibility that Src can induce RGS2 protein activity based on pathway number 7 and given information about each edge. Based on Figure 3, Src induces the expression of RGS2 so maybe the mediators in pathway number 7, has a role in induction of this gene or It might be possible that RGS2 has auto-regulatory effects on its expression.

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 could be a relationship between PNRC1, ETS2 and FATP toward RGS2. All of these relationships are set by PRKCA and PRKG genes at the end of pathways.Nevertheless, PNCR1 makes shorter pathways and is worth more investigation. JUNB and PDP1 each made a relationship with TIAM1. The discovered pathway from JUNB to TIAM1 is mediated by EGFR and SRC demonstrating that SRC could induce its expression by induction of JUNB. Moreover, there are two pathways from TIAM1 to RGS2 and from RGS2 to TIAM1 with the same length which are worth investigating.

bioRxiv preprint first posted online Jan. 29, 2020; doi: http://dx.doi.org/10.1101/2020.01.29.914669. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

References

1. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic acids research, 2002. **30**(1): p. 207-210.
2. Leinonen, R., et al., *The sequence read archive.* Nucleic acids research, 2010. **39**(suppl_1): p. D19-D21.
3. Finn, R., *Targeting Src in breast cancer.* Annals of Oncology, 2008. **19**(8): p. 1379-1386.
4. Gargalionis, A.N., M.V. Karamouzis, and A.G. Papavassiliou, *The molecular rationale of Src inhibition in colorectal carcinomas.* International journal of cancer, 2014. **134**(9): p. 2019-2029.
5. Irby, R.B. and T.J. Yeatman, *Role of Src expression and activation in human cancer.* Oncogene, 2000. **19**(49): p. 5636.
6. Kim, L.C., L. Song, and E.B. Haura, *Src kinases as therapeutic targets for cancer.* Nature reviews Clinical oncology, 2009. **6**(10): p. 587.
7. Gautier, L., et al., *affy—analysis of Affymetrix GeneChip data at the probe level.* Bioinformatics, 2004. **20**(3): p. 307-315.
8. Gentleman, R., et al., *Package 'genefilter'.* 2013.
9. Smyth, G.K., et al., *limma: Linear Models for Microarray and RNA-Seq Data User's Guide.* 2002.
10. Ji, Z., et al., *Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins.* elife, 2015. **4**: p. e08890.
11. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-140.
12. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic acids research, 2000. **28**(1): p. 27-30.
13. Zhang, J.D. and S. Wiemann, *KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor.* Bioinformatics, 2009. **25**(11): p. 1470-1471.
14. Türei, D., T. Korcsmáros, and J. Saez-Rodriguez, *OmniPath: guidelines and gateway for literature-curated signaling pathway resources.* Nature methods, 2016. **13**(12): p. 966.
15. Csardi, G. and T. Nepusz, *The igraph software package for complex network research.* InterJournal, Complex Systems, 2006. **1695**(5): p. 1-9.
16. Oldham, M.C., et al., *Identification and Removal of Outlier Samples Supplement for:" Functional Organization of the Transcriptome in Human Brain.* dim (dat1). **1**(18631): p. 105.
17. Matsuda, M., et al., *abLIM3 is a novel component of adherens junctions with actin-binding activity.* European journal of cell biology, 2010. **89**(11): p. 807-816.

bioRxiv preprint first posted online Jan. 25, 2020; doi: https://doi.org/10.1101/2020.01.25.919639. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

18. Servitja, J.-M., et al., *Rac1 Function Is Required for Src-induced Transformation EVIDENCE OF A ROLE FOR TIAM1 AND VAV2 IN RAC ACTIVATION BY SRC.* Journal of Biological Chemistry, 2003. **278**(36): p. 34339-34346.
19. Bustelo, X.R., *Rac1 function is required for Src-induced transformation: Evidence of a role for Tiam1 and Vav2 in Rac activation by Src.* 2003.
20. Leng, R., et al., *Rac1 expression in epithelial ovarian cancer: effect on cell EMT and clinical outcome.* Medical oncology, 2015. **32**(2): p. 28.
21. Timpson, P., et al., *Coordination of cell polarization and migration by the Rho family GTPases requires Src tyrosine kinase activity.* Current biology, 2001. **11**(23): p. 1836-1846.
22. Mertens, A.E., D.M. Pegtel, and J.G. Collard, *Tiam1 takes PARt in cell polarity.* Trends in cell biology, 2006. **16**(6): p. 308-316.
23. Quarta, S., 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): p. 343-356.
24. Cannito, S., et al., *Hypoxia up-regulates SERPINB3 through HIF-2α in human liver cancer cells.* Oncotarget, 2015. **6**(4): p. 2206.
25. Cannito, S., et al., *SerpinB3 up-regulates hypoxia inducible factors-1α and-2α in liver cancer cells through different mechanisms.* Digestive and Liver Disease, 2016. **48**: p. e19.
26. Cunningham, M.L., et al., *Protein kinase C phosphorylates RGS2 and modulates its capacity for negative regulation of Gα11 signaling.* Journal of Biological Chemistry, 2001. **276**(8): p. 5438-5444.
27. Jiang, Z., et al., *Analysis of RGS2 expression and prognostic significance in stage II and III colorectal cancer.* Bioscience reports, 2010. **30**(6): p. 383-390.
28. Linder, A., et al., *Analysis of regulator of G-protein signalling 2 (RGS2) expression and function during prostate cancer progression.* Scientific reports, 2018. **8**(1): p. 17259.
29. Tan, M., et al., *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): p. 3286-3295.
30. Hou, Y., et al., *Activation of cGMP-dependent protein kinase by protein kinase C.* Journal of Biological Chemistry, 2003. **278**(19): p. 16706-16712.
31. Nalli, A.D., 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): p. 867-880.

bioRxiv preprint first posted online Jan. 5, 2020; doi: http://dx.doi.org/10.1101/2020.01.05.895039. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.