ISTITUTO DI ANALISI DEI SISTEMI ED INFORMATICA
"Antonio Ruberti"
CONSIGLIO NAZIONALE DELLE RICERCHE

F. Conte

AN INTEGRATED NETWORK ANALYSIS FOR UNVEILING CRUCIAL TRANSCRIPTION FACTORS IN TRIPLE-NEGATIVE BREAST CANCER

R. 6, 2018

Federica Conte - Institute for System Analysis and Computer Science "Antonio Ruberti" (IASI), CNR, Via dei Taurini 19, 00185 Rome, Italy. Email: federica.conte@iasi.cnr.it

ISSN: 1128-3378

Collana dei Rapporti dell'Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", CNR

via dei Taurini 19, 00185 ROMA, Italy

tel. ++39-06-49937101/02
fax ++39-06-49937106
email: iasi@iasi.cnr.it
URL: http://www.iasi.cnr.it

Abstract

Among breast cancer subtypes, triple-negative breast cancer (TNBC) is the most aggressive with the highest rates of metastatic disease. TNBC is characterized by the lack expression of estrogen receptor, progesterone receptor, and epidermal growth factor receptor 2, and, consequently, it does not respond to the standard hormonal therapy. Thus, the development of new targeted therapies for TNBC is urgently needed.

To address this issue, here we applied SWIM – a software able to unveil a small pool of genes (called switch genes) associated with drastic changes in cell phenotypes – to gene expression profiles of triple-negative breast invasive carcinoma and matched-normal tissues downloaded from The Cancer Genome Atlas.

SWIM unveiled 293 switch genes, including the transcription factor HMGA1 whose high expression has been recently associated with breast cancer aggressiveness and metastasis.

The oncogenic properties of HMGA1 are mainly due to its capacity to form complexes with other transcription factors that regulate the expression of genes involved in tumor progression and metastasis.

In order to identify new putative transcription factors that could cooperate with HMGA1, we focused on the 12 switch genes that correlated positively with HMGA1 and showed a transcription factor activity.

The activation of HMGA1 and these other 12 TF switch genes could have a key role in the transition from the physiological to pathological state of TNBC subtype.

Key words: triple-negative breast cancer, switch genes, HMGA1

1. Introduction

Breast cancer is the most commonly diagnosed cancer in women and comprises of multiple subtypes with distinct morphologies and clinical implications [1]. Increasing evidence has suggested that breast cancers with different histopathological and biological features exhibit distinct behaviors that lead to different treatment responses. Thus, accurate grouping of breast cancers into clinically relevant subtypes is of particular importance for therapeutic decision.

Despite many factors have been investigated, receptor status has proved to be the most useful in predicting prognosis and responsiveness to treatment of different breast subtype [2]. In particular, breast cancers are classified with respect to the presence or absence of three receptors - i.e. estrogen receptor (ER), progesterone receptor (PR), and epidermal growth factor receptor 2 (HER2) - measured by immunohistochemical techniques.

Breast cancer characterized by the lack expression of these receptors are defined as triple-negative breast cancer (TNBC). TNBC accounts for 15–20% of all invasive breast cancers and has the worst prognosis and the highest rate of metastatic disease compared to other subtypes [3]. At the molecular level, TNBC has significant overlap (approximately 80%) with the basal-like subtype identified on the basis of gene expression profiles [4]. Currently, treatment strategies for TNBC patients are restricted to chemotherapy and then there is an ever-increasing interest for a comprehensive understanding of the molecular basis of TNBC progression in order to identify new efficient targeted therapies.

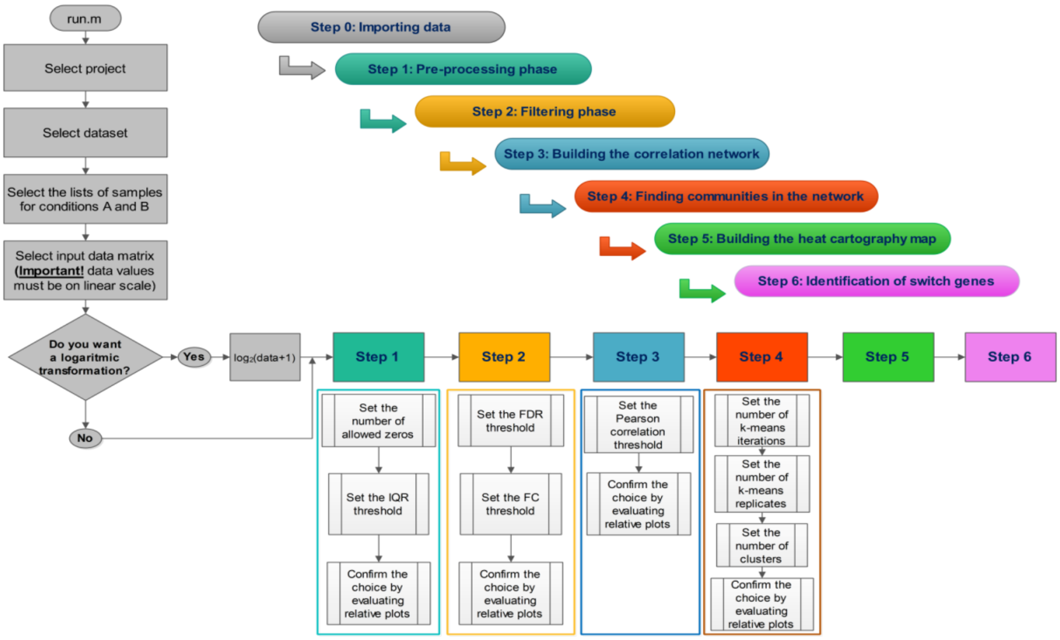
For this aim, we applied the software SWitchMiner (SWIM) [2] - designed to unveil a small pool of genes (called switch genes) associated with drastic changes in cell phenotypes - to gene expression profiles of triple-negative breast invasive carcinoma and matched-normal tissues downloaded from The Cancer Genome Atlas (TCGA).

2. Methods

SWIM software

SWItchMiner (SWIM) is a software developed in MATLAB and downloadable from the supplementary materials of [2]. SWIM implements an integrated network analysis able to extract from genome-wide expression data key players (i.e. switch genes) marking the shift from one condition to another in a complex biological network. The SWIM algorithm steps are shown in Fig.1.

In the present work, SWIM was runned in order to identify switch genes that could be critically associated with the development and aggressiveness of triple-negative breast cancer.



run.m
Select project
Select dataset
Select the lists of samples for conditions A and B
Select input data matrix (Important! data values must be on linear scale)
Do you want a logaritmic transformation?
Yes
No
log₂(data+1)
Step 0: Importing data
Step 1: Pre-processing phase
Step 2: Filtering phase
Step 3: Building the correlation network
Step 4: Finding communities in the network
Step 5: Building the heat cartography map
Step 6: Identification of switch genes
Step 1
Step 2
Step 3
Step 4
Step 5
Step 6
Set the number of allowed zeros
Set the IQR threshold
Confirm the choice by evaluating relative plots
Set the FDR threshold
Set the FC threshold
Confirm the choice by evaluating relative plots
Set the Pearson correlation threshold
Confirm the choice by evaluating relative plots
Set the number of k-means iterations
Set the number of k-means replicates
Set the number of clusters
Confirm the choice by evaluating relative plots

Figure 1 SWIM Flowchart.

Dataset

We analyzed tumour and normal expression data from high-throughput RNA- and miRNA-sequencing of triple-negative breast invasive carcinoma downloaded from the TCGA data portal on 6 December 2014. High-throughput sequencing data for both RNAs and miRNAs correspond to level 3 data (i.e. normalized expression data) given in terms of FPKM (i.e. fragments per kilobase of exon per million fragments mapped). The analysis was restricted to 12 patients for which the complete sets of tumour and matched-normal (i.e., normal tissue taken from the same patient) profiles - for both short and long RNA-seq data - were available.

3. Results

Starting from 20532 RNA-seq profiles and 1047 miRNA-seq profiles, SWIM identified 1655 genes showing significant differential expression (FDR < 0.05) between TNBC and matched-normal tissues (Fig. 2). We found 1217 (74%) differentially expressed genes (DEGs) that were down-regulated in TNBC and the remaining 438 (26%) up-regulated (Fig. 2a).

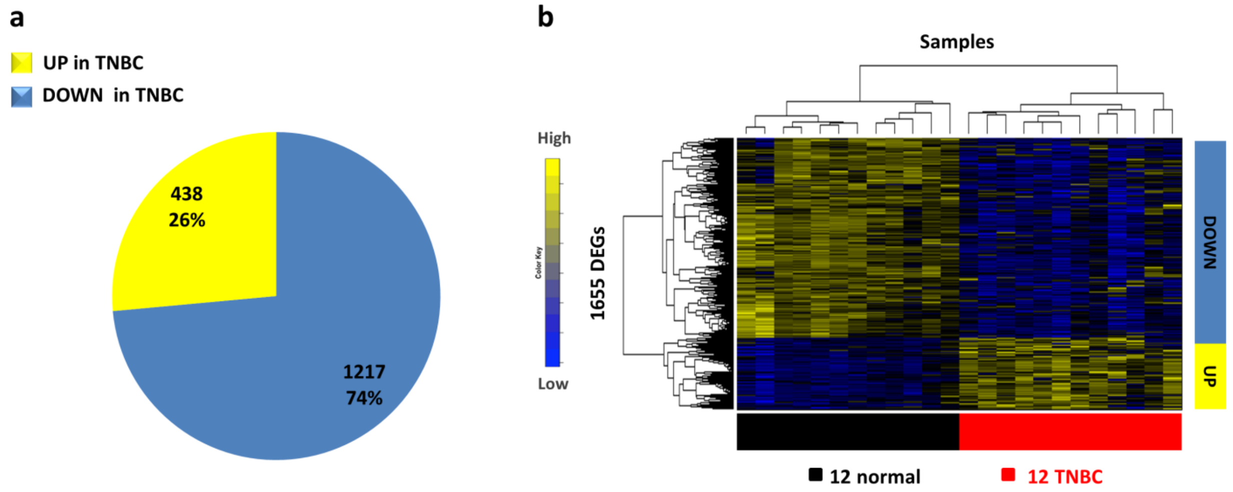


Figure 2. Differentially expressed genes. (a) Pie chart represents the percentages of differentially expressed genes that are up-/down-regulated in TNBC in comparison to normal tissues. (b) Differentially expressed profiles are clustered according to genes (rows) and samples (columns) by using Pearson correlation distance as metrics. Heat map colors represent different expression levels increasing from blue to yellow.

The DEGs matrix of 1655 rows (genes) and 24 columns (12 TNBC samples + 12 matched-normal samples) was used by SWIM to build the correlation network based on the Pearson correlation coefficient between the expression profiles of any pair of the differentially expressed probes (step 3 of Fig.1). In this network, two nodes are connected if the absolute value of the Pearson correlation coefficient for their expression profiles is greater than a given threshold (for this study the selected threshold is 0.7 corresponding to the 70th percentile of the entire correlations' distribution). The correlation network encompasses 1654 nodes (Fig. 3a).

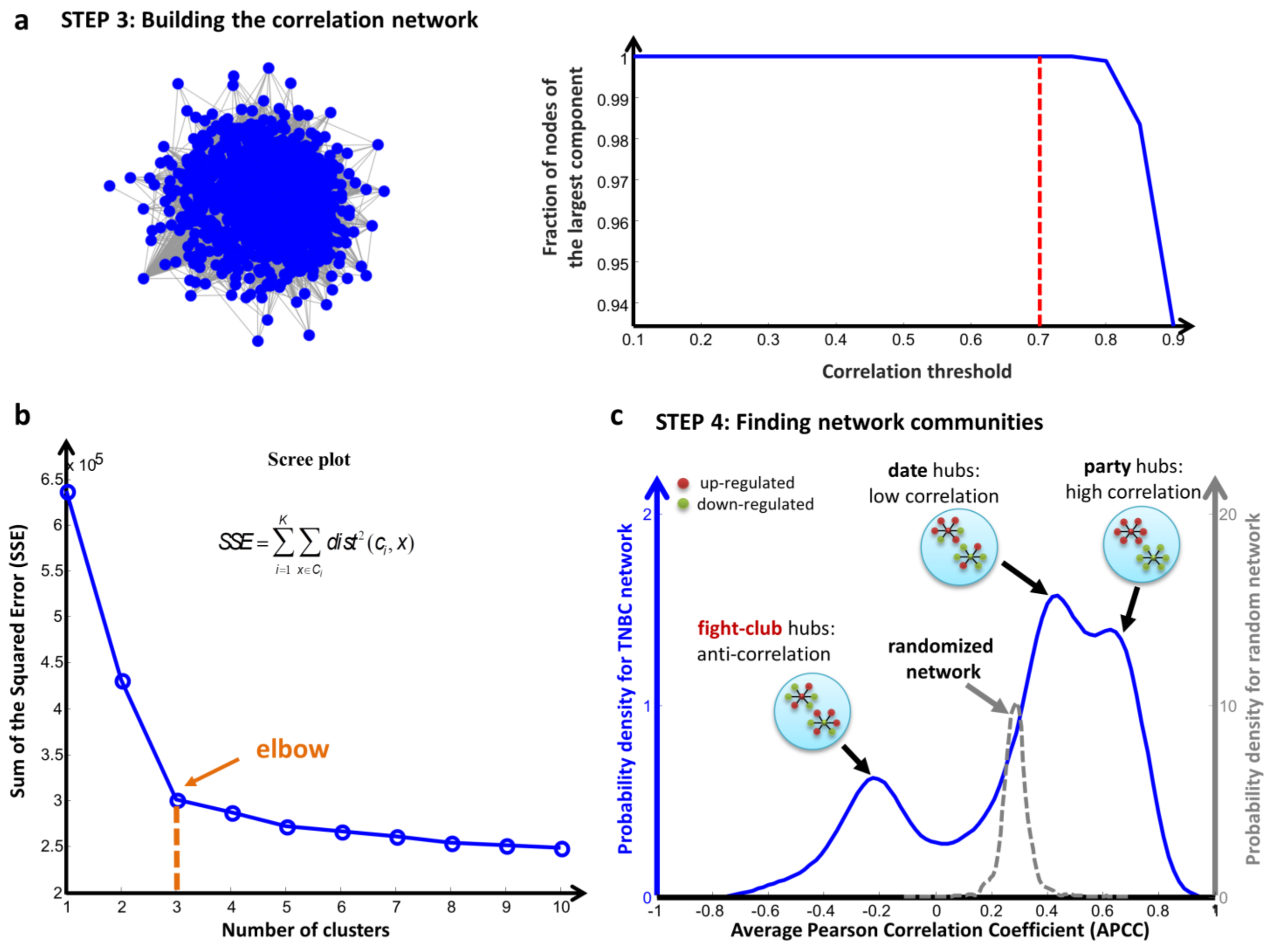


Figure 3. TNBC correlation network and fight-club hubs. (a) [left] Representation of TNBC correlation network; [right] Connectivity of the TNBC correlation network. The x-axis represents the Pearson correlation threshold (ρ = 0.7 or 70th percentile) varying in the chosen range, while the y-axis represents the fraction of nodes populating the largest component. The dashed red lines correspond to the selected threshold. Note that y=1 means that all nodes fall in the largest component and thus the network is fully connected; otherwise more components exist. (b) Scree plot. The position of the elbow suggests a reasonable choice of the number of clusters for the k-means. (c) Probability distribution of APCC for hubs (i.e., node with degree greater than 5) identified in the correlation network built from the TNBC expression dataset (blue solid line) and in its randomized counterpart obtained by shuffling the edges but preserving the degree of each node (grey dashed line).

In order to detect the community structure of the network, SWIM used the k-means clustering algorithm, which partitions n objects (here network nodes) into a predefined number N of clusters. The quality of clustering was evaluated by minimizing the Sum of the Squared Error (SSE), depending on the distance of each object to its closest centroid. As distance measure, SWIM used $dist(x,y) = 1 - \rho(x,y)$, where $\rho(x,y)$ is the Pearson correlation between expression profiles of nodes x and y. A reasonable choice of the number of clusters is suggested by the position of an elbow in the SSE plot computed as function of N (Fig. 3b). The final TNBC correlation network consisted of N=3 clusters.

SWIM next searched for specific topological properties of the correlation network using the date/party/fight-club hub classification system [5], based on the Average Pearson Correlation Coefficients (APCCs) between the expression profiles of each hub (i.e., node with degree greater than 5) and its nearest neighbours (Fig. 3c): date hubs display low co-expression with their partners (i.e. low positive APCC values, APCC < 0.5); party hubs have a high co-expression with their partners (i.e. high positive APCC values, APCC ≥ 0.5);

fight-club hubs show an inversely correlated profile with their partners (i.e. negative APCC values). In TNBC network, SWIM found 346 fight-club hubs, 711 date hubs, and 594 party hubs.

Finally, SWIM assigned a role to each node based on its inter-cluster and intra-cluster interactions. This is reached by defining two statistics: the clusterphobic coefficient $K_{\square}$□and the within-module degree z_g. The clusterphobic coefficient measures the "fear" of being confined in a cluster, in analogy with the claustrophobic disorder. The global within-module degree z_g measures how "well-connected" each node is to other nodes in its own community. According to $K_{\square}$ and z_g values, the plane is divided into seven regions (R1-R7), each defining a specific node role. High z_g values correspond to nodes that are hubs within their module (local hubs), while high values of $K_{\square}$ identify nodes that interact mainly outside their community, i.e. having much more external than internal links. SWIM colored each node in the plane identified by z_g and $K_{\square}$ according to its APCC value, thus defining a heat cartography map (Fig. 4).

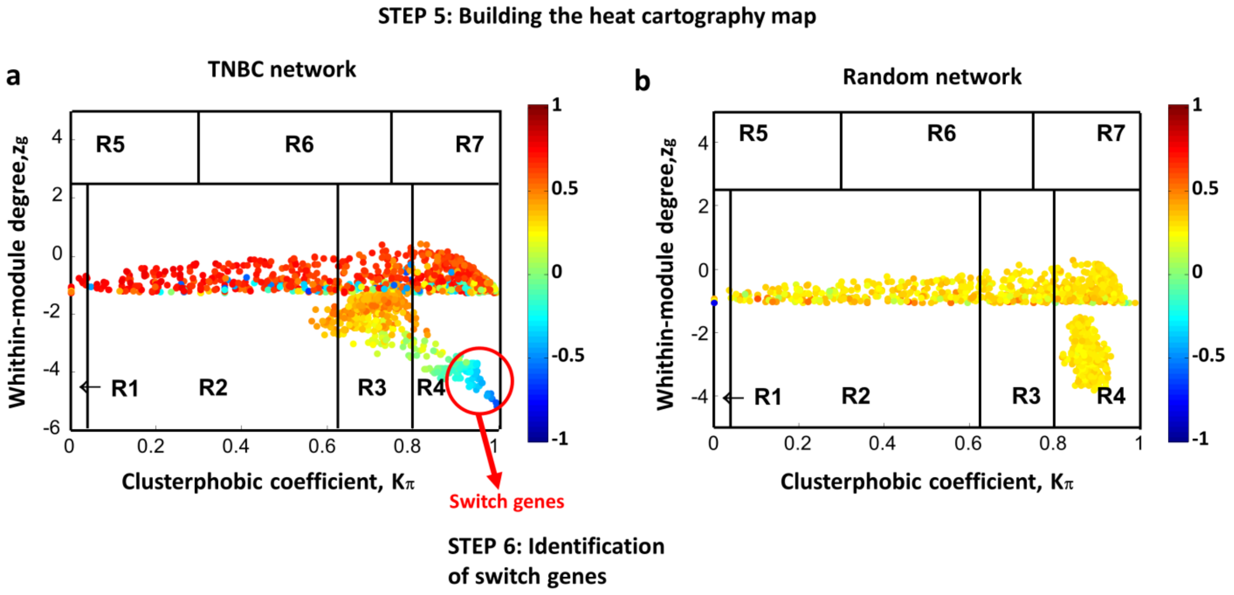
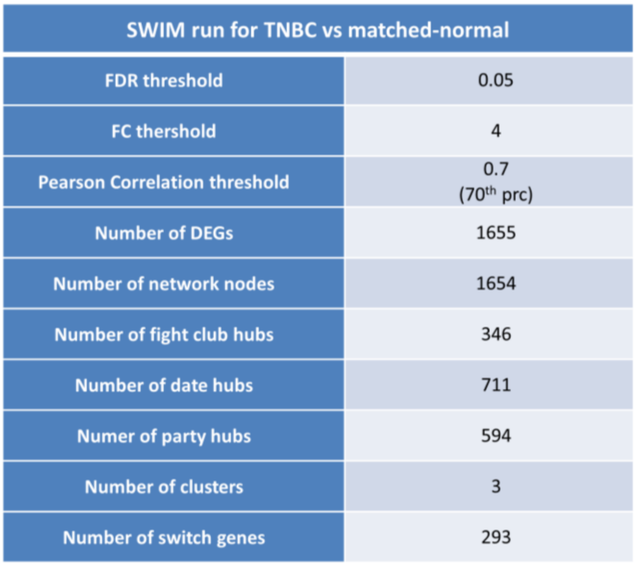


Figure 4. TNBC switch genes. (a-b) Heat cartography maps of TNBC and randomized network. Dots correspond to nodes in the networks. Each node is colored according to the value of the APCC between its expression profile and that of its nearest neighbours in the network.

SWIM identifies "switch genes" as a special subclass of fight-club hubs falling in R4 region. In particular, switch genes have to satisfy the following topological and expression features:

1. being not a hub in their own cluster (z_g<2.5)
2. having many links outside their own cluster ($K_\pi > 0.8$)
3. having a negative average weight of their incident links (APCC < 0)

A summary Table of the SWIM run is provided in Table 1.



SWIM run for TNBC vs matched-normal	
FDR threshold	0.05
FC thershold	4
Pearson Correlation threshold	0.7 (70th prc)
Number of DEGs	1655
Number of network nodes	1654
Number of fight club hubs	346
Number of date hubs	711
Numer of party hubs	594
Number of clusters	3
Number of switch genes	293

Table 5. Summary of SWIM running parameters

We found 293 switch genes in TNBC correlation network. Among them, 275 (94%) are up-regulated in TNBC while the remaining 18 (6%) are down-regulated (Fig. 5a).

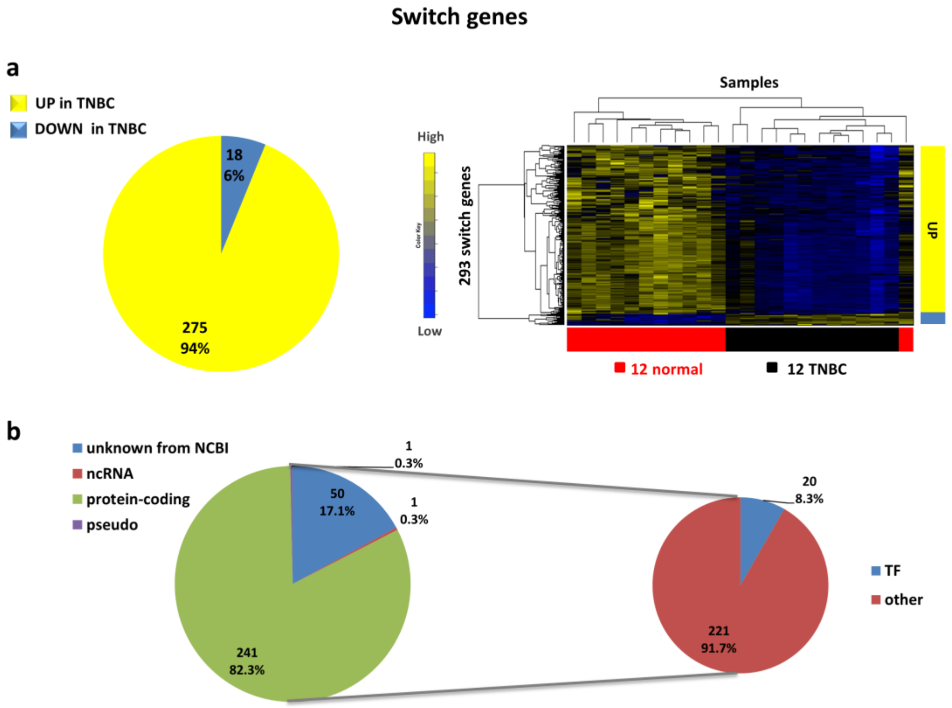
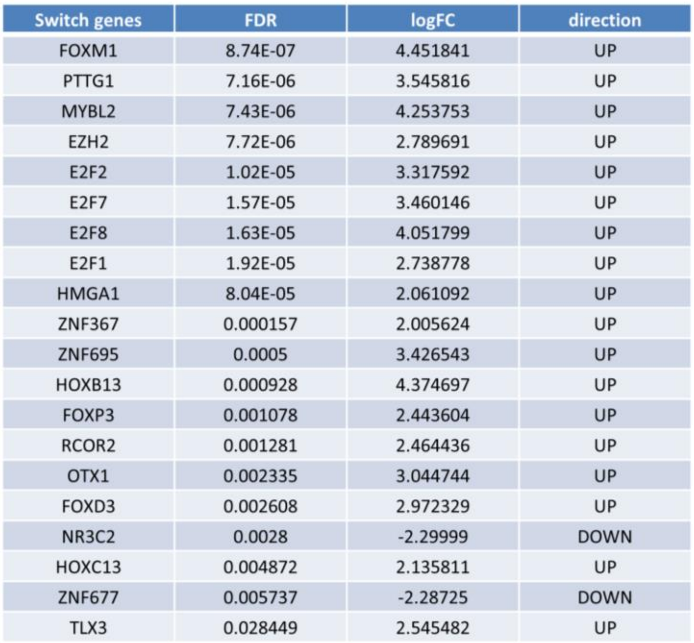


Figure 5. Characterization of TNBC switch genes. (a) [left] Pie chart represents the percentages of switch genes that are up-/down-regulated in TNBC in comparison to normal tissues. [right] Switch genes are clustered according to genes (rows) and samples (columns) by using Pearson correlation distance as metrics. Heat map colors represent different expression levels increasing from blue to yellow. (b) The larger pie chart [left] represents the classification of switch genes according with their molecular type. The smaller pie charts [right] highlight the number of transcription factors among the protein coding switch genes.

The list of switch genes included 1 non-coding RNA, 1 pseudo gene, and 241 protein coding, among which 20 transcription factors (Fig. 5b) that are listed in Table 2.



Switch genes	FDR	logFC	direction
FOXM1	8.74E-07	4.451841	UP
PTTG1	7.16E-06	3.545816	UP
MYBL2	7.43E-06	4.253753	UP
EZH2	7.72E-06	2.789691	UP
E2F2	1.02E-05	3.317592	UP
E2F7	1.57E-05	3.460146	UP
E2F8	1.63E-05	4.051799	UP
E2F1	1.92E-05	2.738778	UP
HMGA1	8.04E-05	2.061092	UP
ZNF367	0.000157	2.005624	UP
ZNF695	0.0005	3.426543	UP
HOXB13	0.000928	4.374697	UP
FOXP3	0.001078	2.443604	UP
RCOR2	0.001281	2.464436	UP
OTX1	0.002335	3.044744	UP
FOXD3	0.002608	2.972329	UP
NR3C2	0.0028	-2.29999	DOWN
HOXC13	0.004872	2.135811	UP
ZNF677	0.005737	-2.28725	DOWN
TLX3	0.028449	2.545482	UP

Table 3. Switch genes that show a transcription factor activity

To provide an overview of the biological functions associated to switch genes, we used FIDEA bioinformatics tool [6]. KEGG pathway analysis revealed that the most significantly over-represented (adjusted p-value < 10^{-3}) pathways among switch genes were ‘Cell cycle’, ‘Oocyte meiosis’, and ‘Progesterone-mediated oocyte maturation’ (Fig. 6).

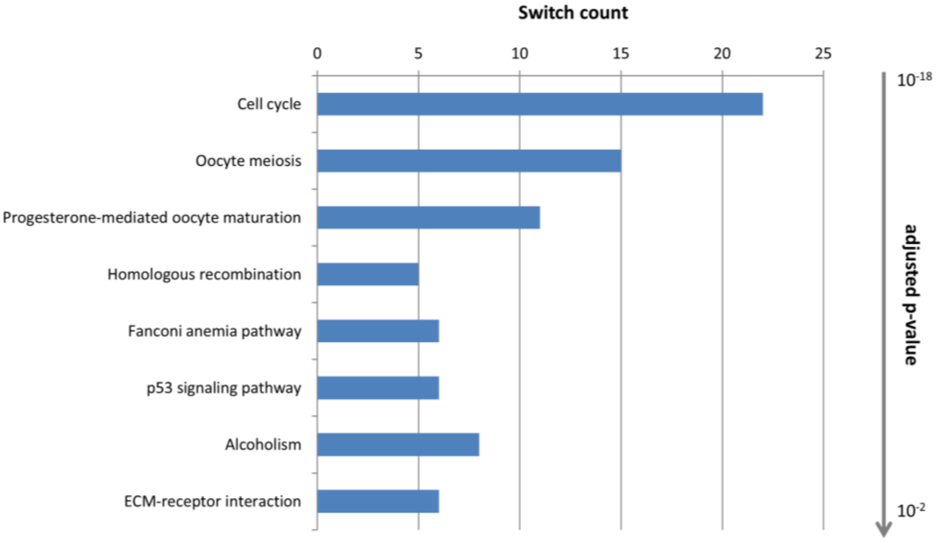


Figure 6. Enriched KEGG pathways among switch genes. Bar plot represents KEGG pathways found to be significantly enriched (adjusted p-value < 0.05) in the lists of switch genes identified by SWIM. KEGG pathways are sorted according to the increasing adjusted p-values.

Interestingly, among switch genes acting as transcription factors (TFs), we found the High Mobility Group A1 (HMGA1) whose high expression has been recently associated with breast cancer aggressiveness and metastasis [7, 8].

The oncogenic activities of HMGA1 are essentially due to its ability to modulate chromatin structure and to form multiple complexes that regulate the expression of genes involved in tumor progression and metastasis [9].

In order to find genes that could cooperate with HMGA1 in controlling the development and aggressiveness of TNBC, we focused on the positive nearest neighbours of HMGA1 (i.e. the set of its adjacent nodes in the correlation network built by SWIM that are highly positively correlated with HMGA1, $\rho > 0.7$).

We found 272 positive nearest neighbours of HMGA1. Among them, we found CCNE2, FOXM1, and E2F family members known to be involved in the regulation of cell cycle progression and proliferation, especially in breast cancer [8, 10, 11]. Interestingly, FOXM1 is also the most significant switch genes acting as TF while E2F family members are in the top-ten (Table 3).

In order to further reduce the list of putative HMGA1-molecular partners in governing breast cancer aggressiveness, we first selected the 221 positive nearest neighbours of HMGA1 acting as switch genes and then, we drew out only those ones showing a transcription factor activity (Fig. 7). In this way, we identified 12 putative TF switch genes cooperating with HMGA1 that would require a subsequent experimental validation.

Finally, since we suppose that switch genes act as negative regulators, we extracted the negative neighbours of HMGA1 (i.e. the set of its adjacent nodes in the correlation network built by SWIM that are highly negatively correlated with HMGA1, $\rho < - 0.7$) and the negative neighbours of the selected 12 TFs. We found a total of 82 negative neighbours shared by these 13 TFs (Fig. 7).

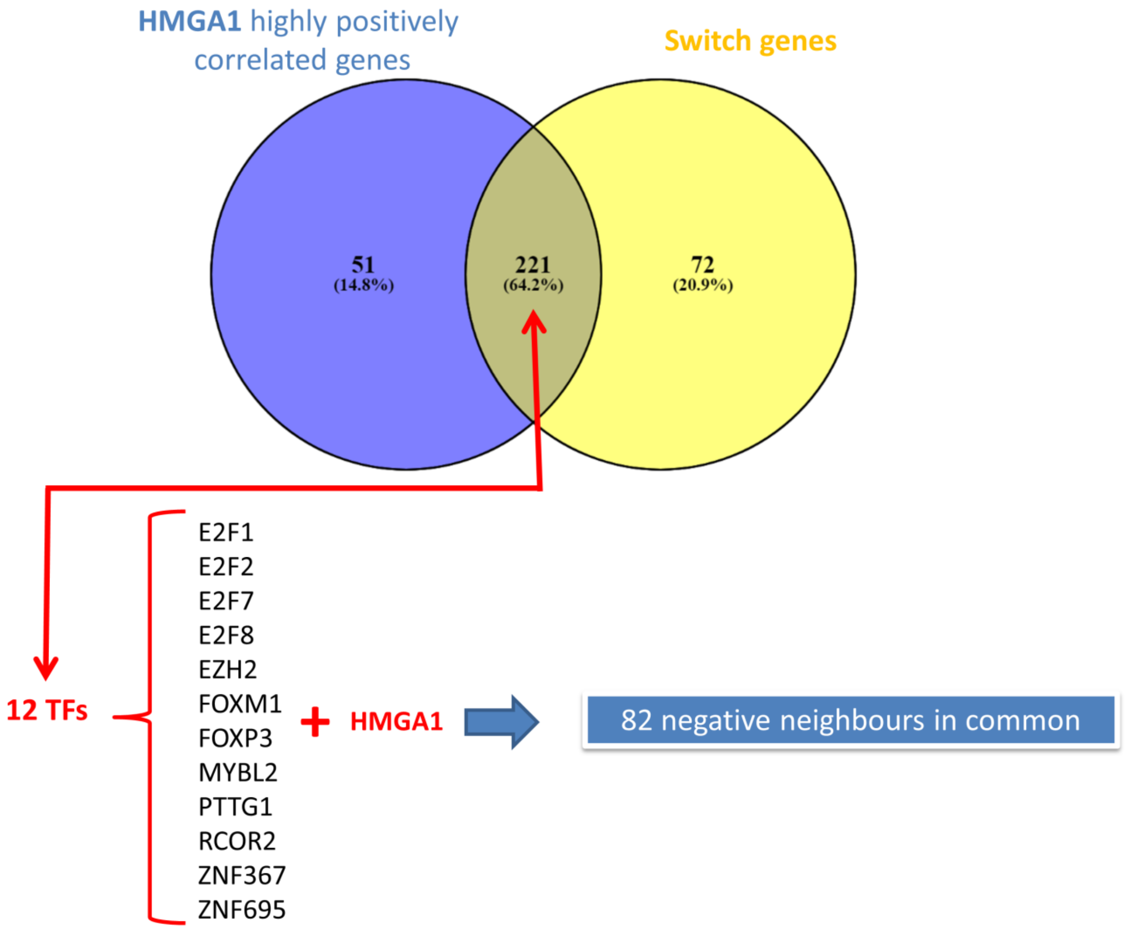


Figure 7. Putative TFs cooperating with HMGA1. Venn diagram between the positive nearest neighbours of HMGA1 and switch genes. Their intersection contains 12 TFs which shared 82 common negative neighbours with HMGA1.

References

[1] Torre, L. A., Siegel, R. L., Ward, E. M., & Jemal, A. (2016). Global cancer incidence and mortality rates and trends—an update. Cancer Epidemiology and Prevention Biomarkers, 25(1), 16-27.

[2] Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., & Shi, B. (2015). Breast cancer intrinsic subtype classification, clinical use and future trends. American journal of cancer research, 5(10), 2929.

[3] Lee, A., & Djamgoz, M. B. (2017). Triple negative breast cancer: emerging therapeutic modalities and novel combination therapies. Cancer treatment reviews.

[4] Foulkes, W. D., Smith, I. E., & Reis-Filho, J. S. (2010). Triple-negative breast cancer. New England journal of medicine, 363(20), 1938-1948.

[5] Paci, P., Colombo, T., Fiscon, G., Gurtner, A., Pavesi, G., & Farina, L. (2017). SWIM: a computational tool to unveiling crucial nodes in complex biological networks. Scientific reports, 7, 44797.

[6] D'andrea, D., Grassi, L., Mazzapioda, M., & Tramontano, A. (2013). FIDEA: a server for the functional interpretation of differential expression analysis. Nucleic acids research, 41(W1), W84-W88.

[7] Pegoraro, S., Ros, G., Piazza, S., Sommaggio, R., Ciani, Y., Rosato, A., ... & Manfioletti, G. (2013). HMGA1 promotes metastatic processes in basal-like breast cancer regulating EMT and stemness. Oncotarget, 4(8), 1293.

[8] Pegoraro, S., Ros, G., Ciani, Y., Sgarra, R., Piazza, S., & Manfioletti, G. (2015). A novel HMGA1-CCNE2-YAP axis regulates breast cancer aggressiveness. Oncotarget, 6(22), 19087.

[9] Sgarra, R., Zammitti, S., Sardo, A. L., Maurizio, E., Arnoldo, L., Pegoraro, S., ... & Manfioletti, G. (2010). HMGA molecular network: from transcriptional regulation to chromatin remodeling. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms, 1799(1), 37-47.

[10] Song, X., Kenston, S. S. F., Zhao, J., Yang, D., & Gu, Y. (2017). Roles of FoxM1 in cell regulation and breast cancer targeting therapy. Medical Oncology, 34(3), 41.

[11] Johnson, J., Thijssen, B., McDermott, U., Garnett, M., Wessels, L. F., & Bernards, R. (2016). Targeting the RB-E2F pathway in breast cancer. Oncogene, 35(37), 4829.