Analysis of transcriptomics in triple-negative breast cancer cells.

Introduction.

Triple negative breast cancer is characterized by the absence of expression of the three genes that are used as biomarkers and drug targets in other types of breast cancer, the estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2).

Aim of this analysis is to identify master regulatory molecules in triple negative breast cancer as well as the regulatory transcription factors that might serve as biomarkers and putative drug targets.

Dataset.

RNA-seq FASTQ files are taken from Gene Expression Omnibus, GSE188914. The reference to the original publication is PMID 35014681. Here, we focus on comparison of gene expression in MDA 231, a triple negative breast cancer cell line, versus MCF-7, an estrogen receptor positive cell line. There are three replicates of each cell line.

Software and databases applied.

The analysis has been done applying the geneXplain platform equipped with the TRANSFAC® and HumanPSD™+TRANSPATH® databases.
Ensembl GRCh38, Entrez, Reactome, HumanCyc databases have been used as they are integrated in the geneXplain platform.

Ready-to-run workflows in the geneXplain platform applied in this study are the following.

- Full RNAseq analysis with subread, featureCounts and limma (single-end)
- Mapping to ontologies (HumanPSD(TM))
- Upstream analysis with feedback loop (TRANSFAC(R) and TRANSPATH(R))

From FASTQ files to DEGs.

Within the geneXplain platform there are ready-to-run workflows (pipelines) which take FASTQ files and output a list of differentially expressed genes (DEGs). The workflows differ by the integrated read alignment and counting methods and by the methods to calculate DEGs. Here, we have applied the workflow with *subread_align* method to map the RNA-seq reads on the reference genome, the *featurecounts* method to count the number of reads and the *limma_voom* method to calculate LogFoldChange and p-value of gene expression.
Three FASTQ files for MDA 231 cells and three FASTQ files for MCF-7 cells have been submitted to the mentioned workflow in the platform.
The workflow output contains a gene table with calculated Log Fold Changes and P-values.

Further, genes were filtered to identify up-regulated, down-regulated as well as non-changed genes.

The details on how the FASTQ files have been uploaded in the geneXplain platform, the workflow structure and run are shown in the video:

https://www.youtube.com/watch?v=linjKZrlRaw&t=1s

The results of the workflow are shown in the video:

https://www.youtube.com/watch?v=rwL-aFYCxbU&t=7s

Functional classification of the up-regulated genes.

The up-regulated genes have been mapped to the Gene Ontology terms, to diseases and to pathways. We aimed to select a functional group of genes for further *Upstream Analysis*.

For this, we applied the next ready-to-run workflow in the geneXplain platform. The HumanPSD and TRANSPATH databases are the integral part of this workflow. The workflow maps the input genes to HumanPSD™ GO categories, TRANSPATH® pathways, Reactome pathways and HumanCyc Pathways, TF classification (TFClass) and also to the HumanPSD™ disease terms.
Let’s consider the results of mapping to the HumanPSD™ disease categories.
Some of the input genes are known biomarkers for particular diseases, this information is manually curated in the HumanPSD™ database. In the column Category, the classification of biomarkers is shown: correlative, causal, and several types more.

For further analysis the 437 genes known to be causal biomarkers for breast neoplasms have been selected. Important to note is that all these genes are significantly up-regulated in triple negative breast cancer cells versus ER-positive breast cancer cells.

The workflow *Functional classification* and its results are shown in details in the video:

https://www.youtube.com/watch?v=6YjLEbdjuQk&t=23s

Upstream analysis.

The workflow *Upstream analysis with feedback loop*s has been applied. The TRANSFAC® and TRANSPATH® databases are integral parts of this workflow. This workflow performs the upstream analysis in two steps. The 1st step is detecting enriched transcription factor binding sites (TFBSs) and the 2nd step identifies the master regulators in the networks upstream of the identified transcription factors. Some master regulators are encoded by the upregulated genes, and they are part of the input set. As these upregulated molecules regulate the network of up-regulated genes, this creates a positive feedback loop.
In the first part of this workflow the input gene set is subjected to the method *Site search on gene set* which predicts TFBSs in the promoters applying TRANSFAC® matrices. In the second part of the workflow the method *Regulator search* is performed on the identified transcription factors using the TRANSPATH® network.

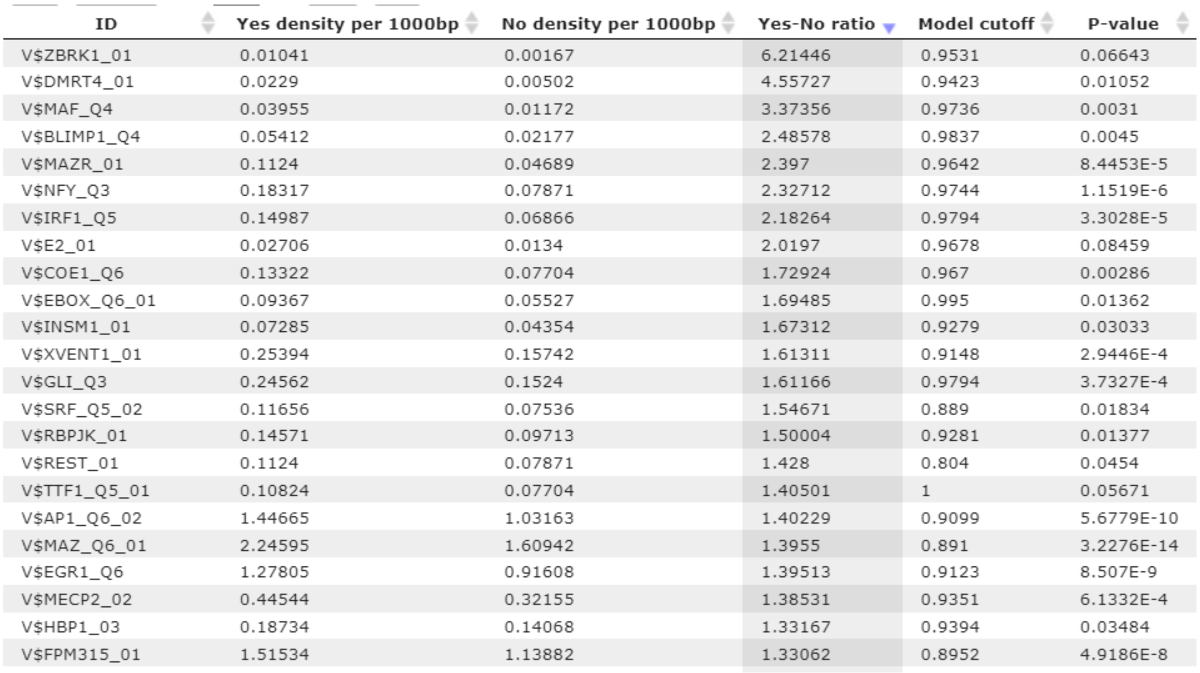
As the input for this workflow, we have taken the 437 genes upregulated in triple negative breast cancer cells and known as causal biomarkers of breast neoplasms. The 546 genes unchanged in the same experiment have been taken as the background set for comparison. The profile with TRANSFAC® matrices and the gene promoter sequences, -1000 to +100 relative to TSS, were taken as per the workflow default.

Promoter analysis and identification of the regulatory transcription factors.

The first part of the workflow has identified transcription factors that might be responsible for regulation of gene expression in breast cancer triple negative cells.

The summary table in the workflow output presents the site models (TRANSFAC matrices) with statistically significant over-represented matches in the promoters of genes up-regulated in triple negative breast cancer cells as compared with ER-positive breast cancer cells. 68 TRANSFAC® matrices with over-represented matches have been identified (**Table 1**).

Table 1. TRANSFAC® matrices with enriched matches in the promoters of triple negative breast cancer cells versus ER-positive breast cancer cells. The enrichment fold of the matches for each matrix (column Yes-No ratio), cutoff value corresponding to the enrichment fold (column Model cutoff) and P-value are shown.



ID	Yes density per 1000bp	No density per 1000bp	Yes-No ratio	Model cutoff	P-value
V$ZBRK1_01	0.01041	0.00167	6.21446	0.9531	0.06643
V$DMRT4_01	0.0229	0.00502	4.55727	0.9423	0.01052
V$MAF_Q4	0.03955	0.01172	3.37356	0.9736	0.0031
V$BLIMP1_Q4	0.05412	0.02177	2.48578	0.9837	0.0045
V$MAZR_01	0.1124	0.04689	2.397	0.9642	8.4453E-5
V$NFY_Q3	0.18317	0.07871	2.32712	0.9744	1.1519E-6
V$IRF1_Q5	0.14987	0.06866	2.18264	0.9794	3.3028E-5
V$E2_01	0.02706	0.0134	2.0197	0.9678	0.08459
V$COE1_Q6	0.13322	0.07704	1.72924	0.967	0.00286
V$EBOX_Q6_01	0.09367	0.05527	1.69485	0.995	0.01362
V$INSM1_01	0.07285	0.04354	1.67312	0.9279	0.03033
V$XVENT1_01	0.25394	0.15742	1.61311	0.9148	2.9446E-4
V$GLI_Q3	0.24562	0.1524	1.61166	0.9794	3.7327E-4
V$SRF_Q5_02	0.11656	0.07536	1.54671	0.889	0.01834
V$RBPJK_01	0.14571	0.09713	1.50004	0.9281	0.01377
V$REST_01	0.1124	0.07871	1.428	0.804	0.0454
V$TTF1_Q5_01	0.10824	0.07704	1.40501	1	0.05671
V$AP1_Q6_02	1.44665	1.03163	1.40229	0.9099	5.6779E-10
V$MAZ_Q6_01	2.24595	1.60942	1.3955	0.891	3.2276E-14
V$EGR1_Q6	1.27805	0.91608	1.39513	0.9123	8.507E-9
V$MECP2_02	0.44544	0.32155	1.38531	0.9351	6.1332E-4
V$HBP1_03	0.18734	0.14068	1.33167	0.9394	0.03484
V$FPM315_01	1.51534	1.13882	1.33062	0.8952	4.9186E-8

These 68 site models have been automatically converted into a list of transcription factors. The conversion resulted in 132 factors. Further, we have checked expression values for these TFs and and found that 46 of these factors are encoded by the upregulated genes (**Table 2**). It is interesting to note that these 46 TFs are upregulated in triple negative breast cancer cells versus ER-positive breast cancer cells, and matches for the corresponding matrices are significantly over-represented in the promoters of upregulated genes.
For example, IRF-1 is upregulated with LogFC equal to 3,0, and the matches of the IRF matrix are over-represented 2,2-fold.

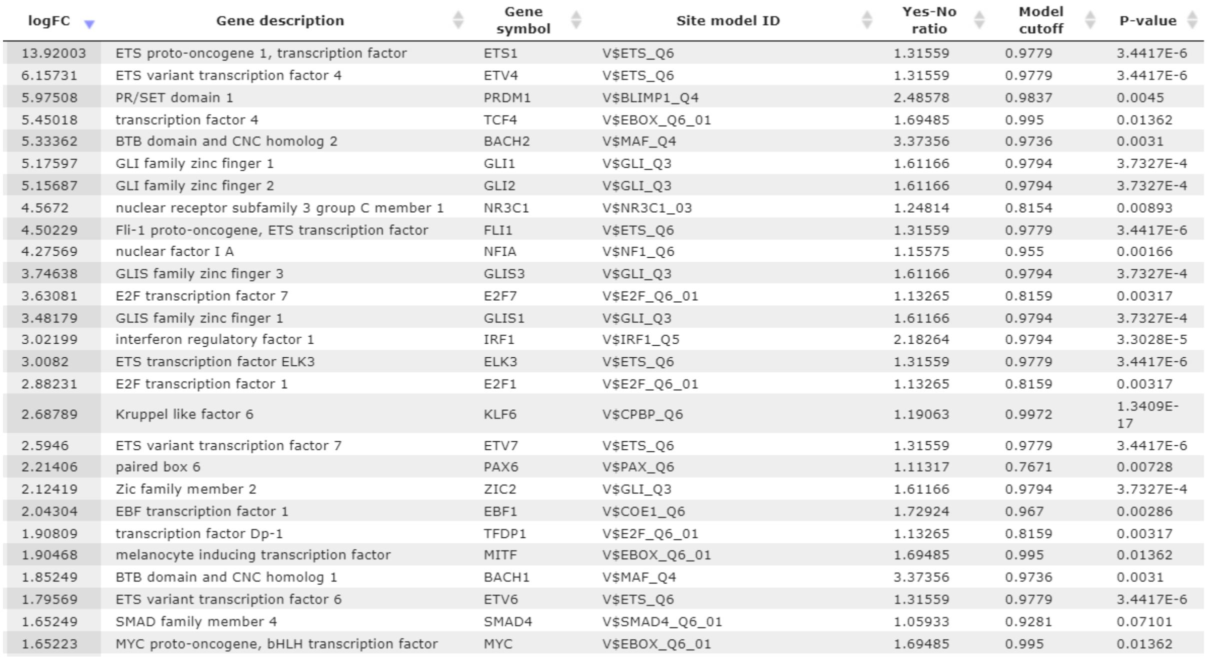
The analysis done suggests that these TFs are responsible for gene regulation in breast cancer triple negative cells. The 132 TFs with the over-represented matches were used as the input for the next step of the workflow, the master regulator search.

Structure of each promoter with the mapped TFBSs can be viewed in details both schematically and on the inbuilt genome browser.

The details of the promoter analysis and visualization of the results are shown in the video

https://www.youtube.com/watch?v=yQV7rWxLSHM

Table 2. Transcription factors resulting from the 1st step of the Upstream Analysis. Each line presents one particular TF. For each TF, gene symbol and gene description are shown ordered by the expression value as LogFoldChange in triple negative breast cancer cells versus ER-positive breast cancer cells (column logFC). Additionally, for each TF there is TRANSFAC® matrix (the column Site model ID), enrichment fold of the matches for this matrix (column Yes-No ratio), cutoff value corresponding to the enrichment fold (column Model cutoff) and P-value.

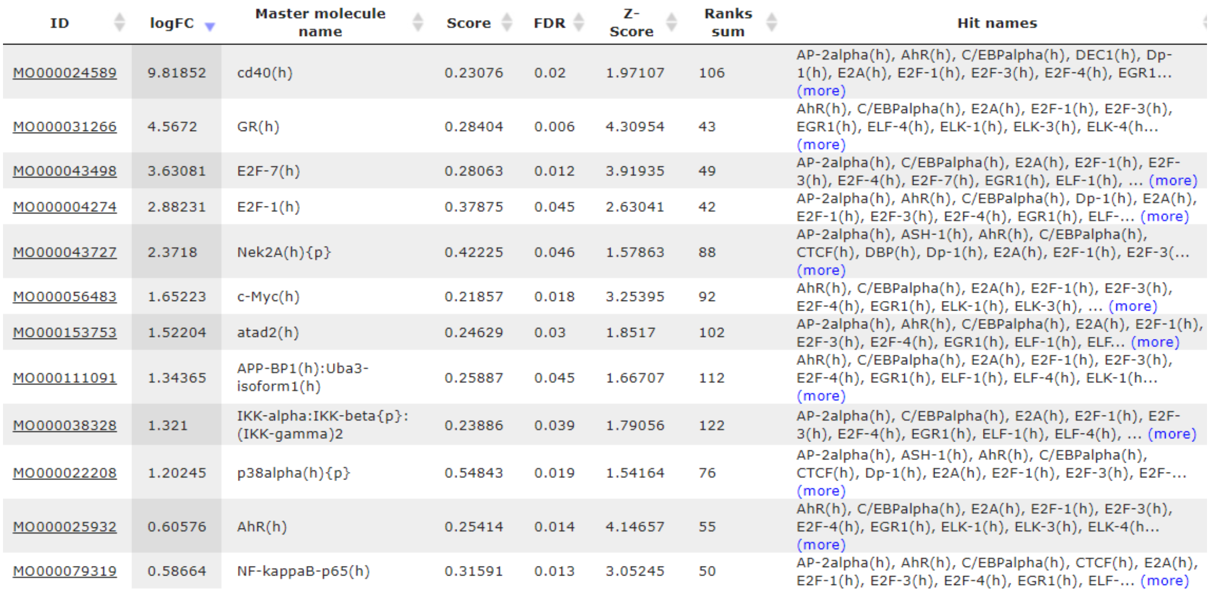


logFC	Gene description	Gene symbol	Site model ID	Yes-No ratio	Model cutoff	P-value
13.92003	ETS proto-oncogene 1, transcription factor	ETS1	V$ETS_Q6	1.31559	0.9779	3.4417E-6
6.15731	ETS variant transcription factor 4	ETV4	V$ETS_Q6	1.31559	0.9779	3.4417E-6
5.97508	PR/SET domain 1	PRDM1	V$BLIMP1_Q4	2.48578	0.9837	0.0045
5.45018	transcription factor 4	TCF4	V$EBOX_Q6_01	1.69485	0.995	0.01362
5.33362	BTB domain and CNC homolog 2	BACH2	V$MAF_Q4	3.37356	0.9736	0.0031
5.17597	GLI family zinc finger 1	GLI1	V$GLI_Q3	1.61166	0.9794	3.7327E-4
5.15687	GLI family zinc finger 2	GLI2	V$GLI_Q3	1.61166	0.9794	3.7327E-4
4.5672	nuclear receptor subfamily 3 group C member 1	NR3C1	V$NR3C1_03	1.24814	0.8154	0.00893
4.50229	Fli-1 proto-oncogene, ETS transcription factor	FLI1	V$ETS_Q6	1.31559	0.9779	3.4417E-6
4.27569	nuclear factor I A	NFIA	V$NF1_Q6	1.15575	0.955	0.00166
3.74638	GLIS family zinc finger 3	GLIS3	V$GLI_Q3	1.61166	0.9794	3.7327E-4
3.63081	E2F transcription factor 7	E2F7	V$E2F_Q6_01	1.13265	0.8159	0.00317
3.48179	GLIS family zinc finger 1	GLIS1	V$GLI_Q3	1.61166	0.9794	3.7327E-4
3.02199	interferon regulatory factor 1	IRF1	V$IRF1_Q5	2.18264	0.9794	3.3028E-5
3.0082	ETS transcription factor ELK3	ELK3	V$ETS_Q6	1.31559	0.9779	3.4417E-6
2.88231	E2F transcription factor 1	E2F1	V$E2F_Q6_01	1.13265	0.8159	0.00317
2.68789	Kruppel like factor 6	KLF6	V$CPBP_Q6	1.19063	0.9972	1.3409E-17
2.5946	ETS variant transcription factor 7	ETV7	V$ETS_Q6	1.31559	0.9779	3.4417E-6
2.21406	paired box 6	PAX6	V$PAX_Q6	1.11317	0.7671	0.00728
2.12419	Zic family member 2	ZIC2	V$GLI_Q3	1.61166	0.9794	3.7327E-4
2.04304	EBF transcription factor 1	EBF1	V$COE1_Q6	1.72924	0.967	0.00286
1.90809	transcription factor Dp-1	TFDP1	V$E2F_Q6_01	1.13265	0.8159	0.00317
1.90468	melanocyte inducing transcription factor	MITF	V$EBOX_Q6_01	1.69485	0.995	0.01362
1.85249	BTB domain and CNC homolog 1	BACH1	V$MAF_Q4	3.37356	0.9736	0.0031
1.79569	ETS variant transcription factor 6	ETV6	V$ETS_Q6	1.31559	0.9779	3.4417E-6
1.65249	SMAD family member 4	SMAD4	V$SMAD4_Q6_01	1.05933	0.9281	0.07101
1.65223	MYC proto-oncogene, bHLH transcription factor	MYC	V$EBOX_Q6_01	1.69485	0.995	0.01362

Identification of the master regulators in networks.

The master regulator search with feedback loops has been performed upstream of the input TFs, which resulted in 12 master molecules (**Table 3**).

Table 3. The resulting list with 12 master regulatory molecules sorted by the LogFC column. Each master regulatory molecule is characterized by a Score, Z-score, FDR, and Ranks Sum. Also, the expression values of the genes encoding master regulators are shown (column logFC). The column “Hit names” shows TFs that can be reached from the corresponding master molecule.



ID	logFC	Master molecule name	Score	FDR	Z-Score	Ranks sum	Hit names
MO000024589	9.81852	cd40(h)	0.23076	0.02	1.97107	106	AP-2alpha(h), AhR(h), C/EBPalpha(h), DEC1(h), Dp-1(h), E2A(h), E2F-1(h), E2F-3(h), E2F-4(h), EGR1... (more)
MO000031266	4.5672	GR(h)	0.28404	0.006	4.30954	43	AhR(h), C/EBPalpha(h), E2A(h), E2F-1(h), E2F-3(h), EGR1(h), ELF-4(h), ELK-1(h), ELK-3(h), ELK-4(h... (more)
MO000043498	3.63081	E2F-7(h)	0.28063	0.012	3.91935	49	AP-2alpha(h), C/EBPalpha(h), E2A(h), E2F-1(h), E2F-3(h), E2F-4(h), E2F-7(h), EGR1(h), ELF-1(h), ... (more)
MO000004274	2.88231	E2F-1(h)	0.37875	0.045	2.63041	42	AP-2alpha(h), AhR(h), C/EBPalpha(h), Dp-1(h), E2A(h), E2F-1(h), E2F-3(h), E2F-4(h), EGR1(h), ELF-... (more)
MO000043727	2.3718	Nek2A(h){p}	0.42225	0.046	1.57863	88	AP-2alpha(h), ASH-1(h), AhR(h), C/EBPalpha(h), CTCF(h), DBP(h), Dp-1(h), E2A(h), E2F-1(h), E2F-3(... (more)
MO000056483	1.65223	c-Myc(h)	0.21857	0.018	3.25395	92	AhR(h), C/EBPalpha(h), E2A(h), E2F-1(h), E2F-3(h), E2F-4(h), EGR1(h), ELK-1(h), ELK-3(h), ... (more)
MO000153753	1.52204	atad2(h)	0.24629	0.03	1.8517	102	AP-2alpha(h), AhR(h), C/EBPalpha(h), E2A(h), E2F-1(h), E2F-3(h), E2F-4(h), EGR1(h), ELF-1(h), ELF... (more)
MO000111091	1.34365	APP-BP1(h):Uba3-isoform1(h)	0.25887	0.045	1.66707	112	AhR(h), C/EBPalpha(h), E2A(h), E2F-1(h), E2F-3(h), E2F-4(h), EGR1(h), ELF-1(h), ELF-4(h), ELK-1(h... (more)
MO000038328	1.321	IKK-alpha:IKK-beta{p}:(IKK-gamma)2	0.23886	0.039	1.79056	122	AP-2alpha(h), C/EBPalpha(h), E2A(h), E2F-1(h), E2F-3(h), E2F-4(h), EGR1(h), ELF-1(h), ELF-4(h), ... (more)
MO000022208	1.20245	p38alpha(h){p}	0.54843	0.019	1.54164	76	AP-2alpha(h), ASH-1(h), AhR(h), C/EBPalpha(h), CTCF(h), Dp-1(h), E2A(h), E2F-1(h), E2F-3(h), E2F-... (more)
MO000025932	0.60576	AhR(h)	0.25414	0.014	4.14657	55	AhR(h), C/EBPalpha(h), E2A(h), E2F-1(h), E2F-3(h), E2F-4(h), EGR1(h), ELK-1(h), ELK-3(h), ELK-4(h... (more)
MO000079319	0.58664	NF-kappaB-p65(h)	0.31591	0.013	3.05245	50	AP-2alpha(h), AhR(h), C/EBPalpha(h), CTCF(h), E2A(h), E2F-1(h), E2F-3(h), E2F-4(h), EGR1(h), ELF-... (more)

These master regulatory molecules are upregulated in triple negative breast cancer cells versus ER-positive breast cancer cells. The top up-regulated molecules are the following: cd40 with LogFC almost 10-fold; GR is upregulated 4,5 times; E2F7 and E2F1 are both upregulated with LogFC 3,6- and 2,9-fold correspondingly; Nek2A with LogFC 2,4-fold; c-Myc with LogFC 1,7-fold. The identified master regulatory molecules are responsible for the coordinated regulation of the networks below, and these networks can explain the differential gene expression in the comparison done. As genes encoding the master molecules are up-regulated by transcription factors in networks downstream of the master molecules, they constitute positive regulatory feedback loops.

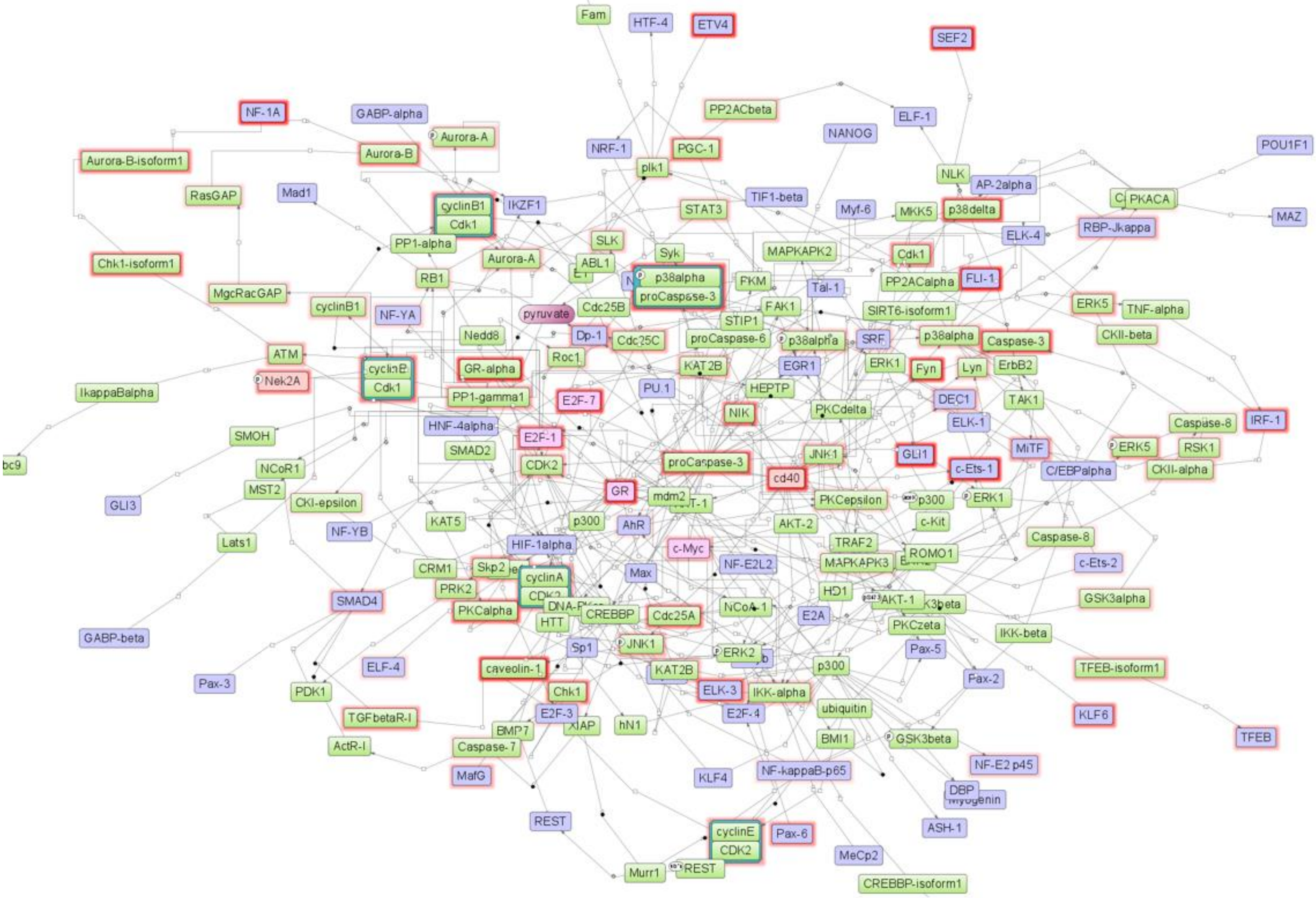
For each individual master regulator or for several master regulators together the identified network can be visualized. As an example, the network of top 6 master molecules is shown (**Fig. 1**)

Our analysis suggests that these molecules play a significant role in regulating pathways in triple negative breast cancer.

The details of the master regulator search and network visualization are shown in the video

https://www.youtube.com/watch?v=e_WFbjpdBe8

Fig 1. The network of top 6 identified master regulators, all of them are encoded by the upregulated genes. The master molecules are shown by the pink color inside their shapes (CD40, GR, E2F1, E2F7, Nek1A and c-Myc). Blue color shows TFs that were used as input into the master regulator search, and green color shows intermediate molecules. All the reactions between molecules are integrated from the TRANSPATH® database. The red frames highlight upregulated molecules, and the intensity of the frame color reflects the fold of upregulation.



Discussion.

The analysis presented has identified promising master regulators in the signaling networks in triple negative breast cancer cells.

CD40 (in our diagram: cd-40) is a member of the TNF receptor family, and its association with triple-negative tumor has been published (PMID: **32256143**). According to the expression data analyzed here, the gene encoding CD40 is highly over-expressed in MDA 231, a triple negative breast cancer cell line, versus MCF-7, an estrogen receptor positive cell line. The LogFoldChange for CD40 gene is 9.8 and it is the top up-regulated master regulatory molecule (**Table 3**).

As we know from the HumanPSD™ database, CD40 is associated with 70 diseases as a biomarker, among them a number of different tumors (Locus Report for CD40, https://portal.genexplain.com/cgibin/build_hpt/idb/1.0/pageview.cgi?view=LocusReport&gene_acc=GN000003091). Association of CD40 with breast neoplasms in general is well documented.

Here we show that very high expression levels of CD40 might be a specific marker of triple negative breast cancer cells even as compared to ER-positive breast cancer cells. Our study suggests a master regulatory role of CD40 specifically in triple negative breast cancer cells.