Original Article

Identification of novel prognostic genes of triple-negative breast cancer using meta-analysis and weighted gene co-expressed network analysis

Wenning Cao[1,2], Yike Jiang[3,4], Xiang Ji[2,5], Xuejiao Guan[2,5], Qianyu Lin[2,4], Lan Ma[2,3,4,6]

[1]Department of Chemistry, Tsinghua University, Beijing, China; [2]State Key Laboratory of Chemical Oncogenomics, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China; [3]Institute of Biomedical Health Technology and Engineering, Shenzhen Bay Laboratory, Shenzhen, China; [4]Precision Medicine and Healthcare Research Center, Tsinghua-Berkeley Shenzhen Institute, Shenzhen, China; [5]School of Life Science, Tsinghua University, Beijing, China; [6]Institute of Biopharmaceutical and Health Engineering, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China

Contributions: (I) Conception and design: W Cao, L Ma; (II) Administrative support: Y Jiang, M Lan; (III) Provision of study materials or patients: None; (IV) Collection and assembly of data: W Cao, X Guan; (V) Data analysis and interpretation: W Cao, Y Jiang, X Ji, Q Lin; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Prof. Lan Ma. Institute of Biopharmaceutical and Health Engineering, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China. Email: malan@sz.tsinghua.edu.cn.

Background: Triple-negative breast cancer (TNBC) is an aggressive subtype of breast cancer with high rates of metastasis and recurrence. Conventional clinical treatments are ineffective for it as it lacks therapeutic biomarkers. Figuring out the biomarkers related to TNBC will be beneficial for its clinical treatment and prognosis.

Methods: Five independent datasets downloaded from the Gene Expression Omnibus database were merged to identify differentially expressed genes between TNBC and non-TNBC samples by using the MetaDE.ES method followed by mapping the differentially expressed genes into a protein-protein interaction network. Meanwhile, the weighted gene co-expressed network analysis (WGCNA) of The Cancer Genome Atlas data was performed to screen the hub genes. The gene functional analyses were conducted by Gene Ontology (GO) enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis. The correlation between gene expression level and patient overall survival was evaluated by survival analysis.

Results: A total of 11 differentially expressed genes (*CDH1*, *SP1*, *MYC*, *FAF2*, *IFI16*, *MDM2*, *AR*, *DBN1*, *HSPB1*, *FLNA*, *YWHAB*) were obtained from the protein-protein interaction network with degree >10. WGCNA revealed 5 hub genes (*TPX2*, *CTPS1*, *KIF2C*, *MELK*, *CDCA8*) that were significantly associated with TNBC. Cell cycle, oocyte meiosis, spliceosome were the pathways significantly enriched in these genes according to GO functionally annotated terms and KEGG pathways analysis. The Kaplan-Meier curves showed that the expression levels of *HSPB1*, *IFI16*, *TPX2* were significantly associated with the survival time of TNBC patients ($P<0.05$).

Conclusions: A total of 16 genes significantly associated with TNBC were identified by bioinformatic analyses. Among these 16 genes, *HSPB1*, *IFI16*, *TPX2* might be able to be used as biomarkers of TNBC.

Keywords: Triple-negative breast cancer (TNBC); meta-analysis; weighted gene co-expressed network analysis (WGCNA); differentially expressed genes; prognostic biomarkers

Submitted Aug 20, 2020. Accepted for publication Nov 08, 2020.
doi: 10.21037/atm-20-5989
View this article at: http://dx.doi.org/10.21037/atm-20-5989

Introduction

Triple-negative breast cancer (TNBC) is a subtype of breast cancers that lacks expression of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor type 2 (HER2) (1). Although TNBC only accounts for about 15–20% of all breast cancers, it is more aggressive than other subtypes of breast cancers (2). Moreover, TNBC patients have high rates of metastasis and recurrence, especially within the first 5 years after diagnosis (1,2). All these unfavorable factors collectively lead to poor prognosis of TNBC patients. Unfortunately, due to lack of expression of hormone receptors (ER and PR) and HER2, conventional hormone therapy (e.g., tamoxifen) and anti-HER2 antibody therapy (e.g., trastuzumab) are ineffective for TNBC. Surgery combined with chemotherapy and radiotherapy is still the most commonly used clinical therapeutic strategy for TNBC (3). Hence, it is urgent to identify biomarkers of TNBC, which would be beneficial for TNBC early detection, prediction of prognosis, and development of TNBC targeted drugs.

With the prevalence of microarray technologies, efforts were devoted to identifying the TNBC-sensitive and specific signatures (4,5). Although microarray-based studies are informative, some studies reported that the results of single microarray analyses were not reproducible or not robust (6,7). Meta-analysis is a statistical technique that includes multiple studies to yield a more precise and reliable evaluation of differentially expressed genes (DEGs) (8). The weighted gene co-expressed network analysis (WGCNA) has been proven as an effective method to group genes that have similar expression patterns into a model related to the desired traits (9). These two bioinformatics methods have been commonly applied to identify interested gene sets.

In this study, an integrated analysis was applied to 5 independent datasets of TNBC and non-TNBC samples, which identified 361 dysregulated genes. The protein and protein interaction (PPI) network was subsequently employed to seek the hub genes. WGCNA was used to seek highly correlated genes among modules that correlated to TNBC, which may share important biological regulatory roles. A panoramic view of molecular mechanisms related to TNBC was obtained through using a series of functional annotations including Gene Ontology (GO) analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis. Finally, identification of potential prognostic molecules was achieved based on patients' clinical survival data and candidate genes in both analyses. Our study not only provides promising therapeutic and prognostic targets of TNBC, but also promotes understanding of the molecular mechanisms of TNBC. We present the following article in accordance with the MDAR reporting checklist (available at http://dx.doi.org/10.21037/atm-20-5989).

Methods

Identification of DEGs and hub genes in PPI network

Gene expression profile datasets that met the inclusion criteria were retrieved from the Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) database. The inclusion criteria included: (I) mRNA expression profiling by array of homo sapiens; (II) tissue samples from TNBC and other kinds of breast cancer; (III) number of samples >15. Five microarray datasets containing 63 TNBC samples and 169 non-TNBC samples were collected (*Table 1*), including GSE27447, GSE36295, GSE61724, GSE43358, and GSE75678. The raw data (CEL files) of the first four datasets were downloaded followed by background correction, normalization, and median polish summarization with the Robust Multichip Average algorithm analysis (10). The normalized data (txt files) of the dataset GSE75678 was also downloaded. The probe IDs were converted to official gene symbols according to the annotation files. The probes without mapped genes were discarded while the average expression value of the multiple probes mapped to the identical gene was obtained.

After merging 5 datasets and filtering non-expressed and non-informative genes, quality control (QC) was carried out to determine whether a study should be included or excluded by using the MetaQC package (11), which provided six QC measurements: (I) internal homogeneity of co-expression structure among studies (IQC); (II) external consistency of co-expression structure correlating with a pathway database (EQC); (III) accuracy of DEG detection (AQCg) or pathway identification (AQCp); (IV) consistency of differential expression ranking of genes (CQCg) or pathways (CQCp). The MetaDE.ES method was adopted to identify DEGs (12). Firstly, the heterogeneity test which was used to determine gene expression differentiation among various datasets indicated no heterogeneity with the thresholds: Q P value (Qpval) >0.05 and τ^2=0. Secondly, genes with a false discovery rate (FDR) <0.01 were considered as DEGs between TNBC and non-TNBC.

The gene interactions of humans in the Human Protein Reference Database (www.hprd.org/, HPRD Release 9) and Biological General Repository for Interaction Datasets (thebiogrid.org/, BioGRID Version 3.5.165)

Table 1 Characteristics of individual studies included in meta-analysis

GEO accession	Chip	Number of probes	Sample size	Number of TNBC	Number of non-TNBC
GSE27447	[HuGene-1_0-st] Affymetrix Human Gene 1.0 ST Array	33297	19	5	14
GSE43358	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	54675	57	17	40
GSE36295	[HuGene-1_0-st] Affymetrix Human Gene 1.0 ST Array	33297	38	11	27
GSE61724	[HuGene-1_0-st] Affymetrix Human Gene 1.0 ST Array	33297	64	16	48
GSE75678	Agilent-014850 Whole Human Genome Microarray	45220	54	14	40
		Total	232	63	169

GEO, gene expression omnibus; TNBC, triple-negative breast cancer.

were downloaded (13,14). The interactions of identified DEGs based on these two databases were subjected to Cytoscape 3.6.1 for visualization (15). In the PPI network, "nodes" represented proteins, and "edges" represented the interactions between two proteins. "Degree" described the quantitative relationship of edges between different nodes. That is, a gene with more degrees implies more important roles in biological processes.

WGCNA analysis of hub genes

Firstly, Gene expression data of all breast cancer samples were downloaded from The Cancer Genome Atlas (TCGA; https://cancergenome.nih.gov/). TCGA level 3 expression data of breast cancer and corresponding trait information including 1,090 samples (https://cdn.amegroups.cn/static/public/10.21037atm-20-5989-1.docx) were used as input. Outlier samples were removed, and missing values were filtered. A matrix of similarity was constructed according to Pearson's correlation coefficient among all genes. The soft threshold parameter which satisfied the scale-free co-expression network relationship was set to 4. Secondly, the topological overlap matrix was transformed from the adjacency matrix, and the corresponding dissimilarity was calculating to identity hierarchical clustering genes through the dynamics cut tree algorithms. Different modules eigenvectors (MEs) were achieved to merge modules with high similarity at the height cut of 0.75. The relevance between modules and clinical traits (TNBC and other molecular subtypes of breast cancers) was shown with heatmap to figure out modules most closely associated with TNBC. Gene significance (GS) was defined to measure the correlation between the gene and the trait, and another term named module membership (MM) was used to quantify the correlation of the MEs and the gene expression profile. Hub genes in the key modules were defined by high MM and GS, i.e., genes in the key modules with MM >0.8 and GS >0.2 were selected for further analysis. WGCNA was implemented by the WGCNA package in R (16).

Functional analysis of gene sets

To investigate the biological roles of gene sets obtained from meta-analysis and WGCNA in TNBC, KEGG pathway enrichment, and GO enrichment analysis which represented 3 categories including Biological Process (BP), Molecular Function (MF), and Cellular Component (CC) were performed. GO analysis and pathway analysis of DEGs were based on online website DAVID (https://david.ncifcrf.gov/) and KOBAS 3.0 (http://kobas.cbi.pku.edu.cn/), respectively. Additionally, functional analysis of key modules was performed by using the clusterProfiler package in R (17).

Survival analysis

To screen prognostic signatures, the critical genes obtained from meta-analysis and WGCNA were selected for survival analysis, which was performed based on TCGA gene expression and clinical information. Kaplan-Meier survival curves were plotted by the "survival" package in R to show the relationship between overall survival (OS) and gene expression level. The log-rank test was used to generate P values. Genes with a threshold of $P<0.05$ were considered to have a significant difference during patients' survival time.

Statistical analysis

All analyses were performed using R (version 3.6),

Table 2 MetaQC quantitative quality control measures for selected datasets.

Dataset	IQC	EQC	CQCg	CQCp	AQCg	AQCp	Rank
GSE43353	4.91	4	59.48	23.64	41.5	25.54	2.00
GSE75678	0.61*	3.82	97.4	62.92	68.53	36.21	2.08
GSE27447	3.14	4	17.81	3.39	11.45	8.21	3.42
GSE36295	4.91	3.82	14.14	17.3	5.53	10.24	3.67
GSE61724	5.31	3.7	15.31	7.95	8.53	5.38	3.83

*, low performance. IQC, internal quality control; EQC, external quality control; CQCg, consistency quality control of differential expression ranking in genes; CQCp, consistency quality control of differential expression ranking in pathways; AQCg, accuracy quality control of differentially expressed gene detection; AQCp, accuracy quality control of pathway identification.

online website DAVID and KOBAS 3.0. The DEGs were identified by combining effect sizes in the meta-analysis. The functional analysis was conducted using the hypergeometric test. The survival analysis was performed by the log-rank test. P<0.05 was considered as statistically significant in all analyses except for the meta-analysis, which adopted FDR <0.01.

Results

Quality assessment and DEGs screening

The results of QC are listed in *Table 2*. All datasets were included for further analysis as all datasets satisfied the selection criteria, i.e., datasets with good performance in at least 5 QC measures. Using the metaDE.ES method, 361 DEGs between TNBC and non-TNBC samples were identified with the screening criteria: Qpval >0.05, τ^2=0, and FDR <0.01, including 146 up-regulated genes and 215 down-regulated genes (https://cdn.amegroups.cn/static/public/10.21037atm-20-5989-2.docx).

Hub genes in PPI network

After mapping 361 DEGs into the PPI database, the PPI network comprised of 222 nodes and 375 edges was constructed (*Figure 1*). According to network analysis, 11 genes with degree >10 were listed in *Table 3*.

TNBC-associated modules and hub genes in WGCNA

As shown in *Figure 2A*, 16 modules were identified by average linkage hierarchical clustering based on expression values of 19,641 genes. Genes in the grey module were excluded as it contained genes that were not able to be clustered into other modules. At the height cut of 0.75, the blue module and pink module were grouped into the blue module. The black module and magenta module were grouped into the black module. As a result, 14 modules were achieved eventually. Next, we correlated genes and modules with clinical features. Based on the correlation between ME and clinical traits shown in the heatmap (*Figure 2B*), we found cyan module (r=0.64, P=$1e^{-121}$) and yellow module (r=0.56, P=$7e^{-85}$) were significantly positively associated with TNBC. Apart from this, the correlation between ME and ER^-, PR^-, $HER2^-$-breast cancers roughly showed a consistent tendency with TNBC. Comparing GS in each module, the results showed that cyan and yellow modules had the strongest correlation with TNBC (*Figure 2C*). The cluster analysis and heatmap shown in *Figure 2D* further strengthened the evidence that these two modules were highly related to TNBC. For each gene in these two modules, GS, MM, and intramodule connectivity (KME) were calculated to draw scatterplots, respectively. It was obvious that GS was highly positively connected with MM in the cyan module (*Figure 2E*), while genes with high GS had relative lower KME in the cyan module (*Figure 2F*). GS was also highly positively correlated to MM in the yellow module (*Figure 2G*), but genes with high GS had higher KME in the yellow module (*Figure 2H*). Hub genes in the yellow modules were *TPX2*, *CTPS1*, *KIF2C*, *MELK*, and *CDCA8* (*Figure 2I*), while no hub genes with GS >0.2 and MM >0.8 were found in the cyan module.

Functional analysis of interesting gene sets

KEGG pathway enrichment and GO enrichment were performed for DEGs and genes in TNBC-associated modules (*Figure 3*, https://cdn.amegroups.cn/static/public/10.21037atm-20-5989-3.docx, https://cdn.

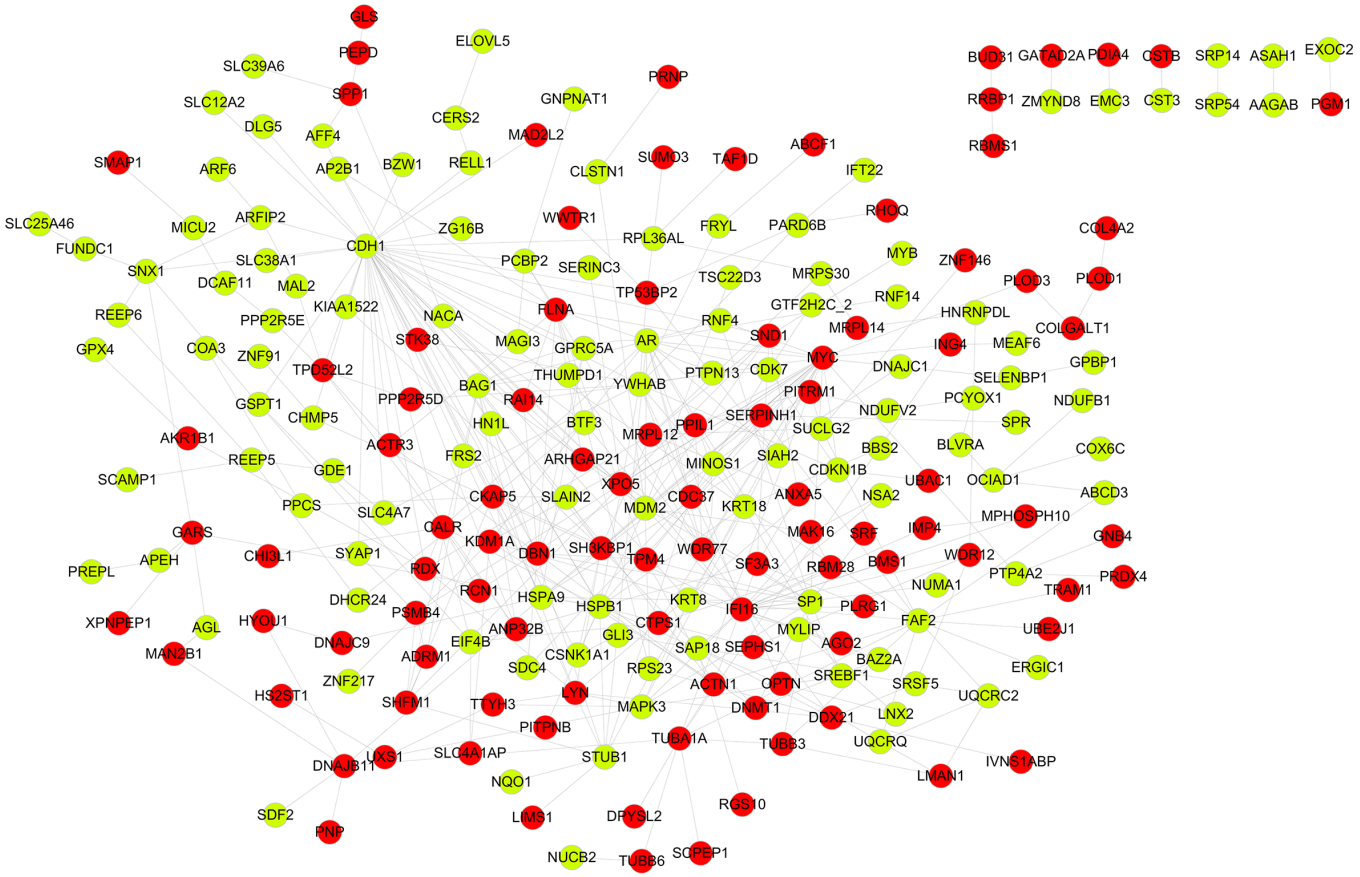

Figure 1 PPI network of DEGs. Red nodes represent higher expression in TNBC samples than non-TNBC samples, while green nodes represent lower expression in TNBC samples than non-TNBC sample. PPI, protein-protein interaction; DEGs, differentially expressed genes; TNBC, triple-negative breast cancer.

Table 3 Top 11 genes with degree >10 in PPI network

Gene	Expression	Degree	Q.value	Qpval	τ^2	FDR
CDH1	Down	42	1.33884	0.85475	0	0.0053
MYC	Up	23	2.55752	0.63437	0	0.0006
IFI16	Up	18	3.96701	0.41049	0	8.89E-03
AR	Down	18	1.67509	0.79524	0	3.25E-19
HSPB1	Down	15	3.70639	0.44720	0	3.25E-19
YWHAB	Down	15	0.47565	0.97583	0	0.0074
SP1	Down	14	2.23518	0.69259	0	0.0029
FAF2	Down	13	2.37068	0.66793	0	0.0011
MDM2	Down	13	2.95618	0.56519	0	0.0099
DBN1	Up	11	1.77908	0.77631	0	0.0023
FLNA	Up	11	2.86412	0.58082	0	0.0023

PPI, protein interaction; Qpval: Q P value; FDR: false discovery rate.

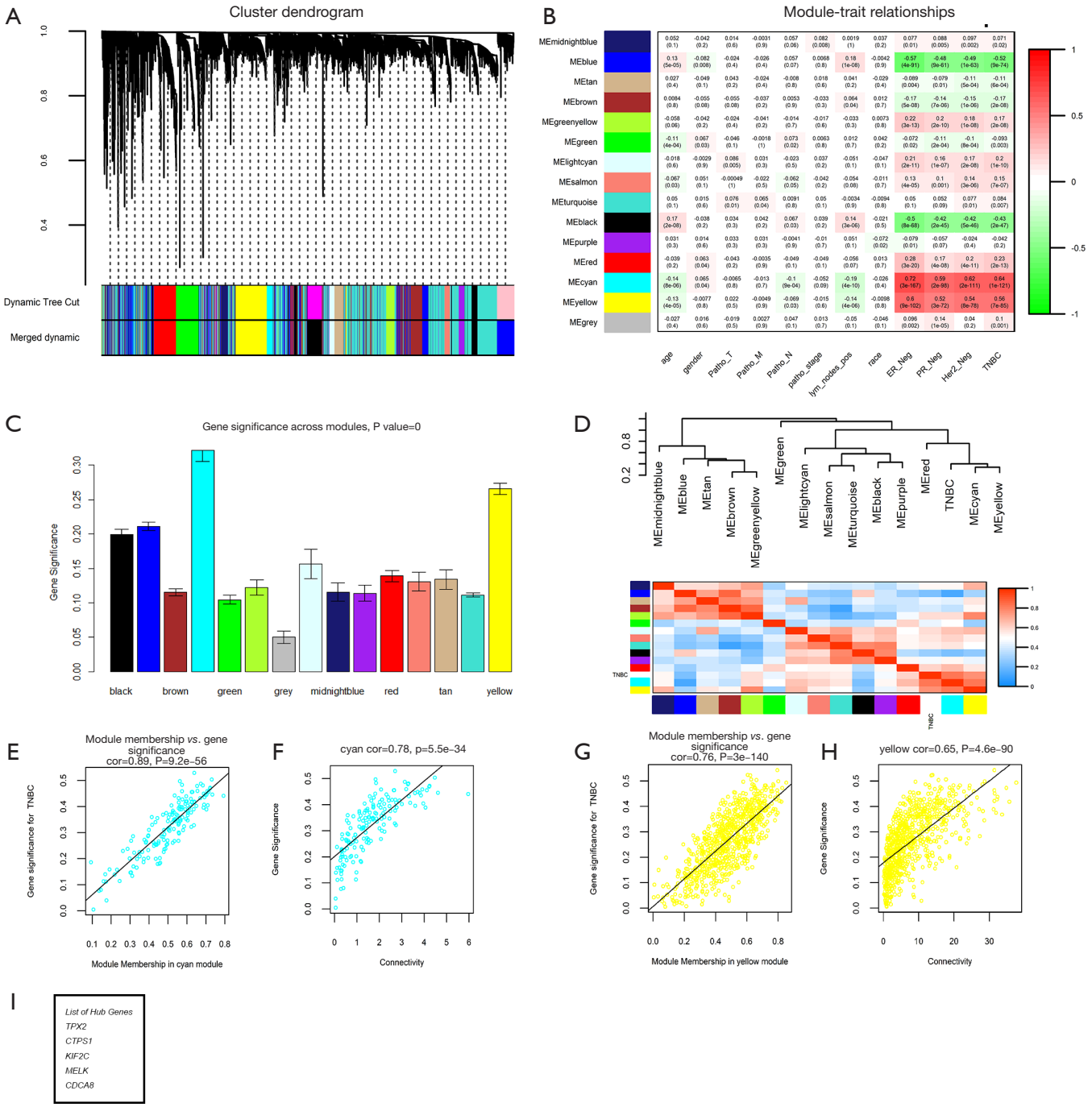

Figure 2 The process of WGCNA analysis. (A) Hierarchical clustering dendrograms of identified co-expressed genes in modules. (B) Module-trait relationships. The Y axis is the modules which are represented by "ME + color". The X axis is the clinical features. Patho_T, Patho_M, and Patho_N stand for the TNM system. Patho_stage is stage of cancer. Lym_nodes_pos means lymph node positive. ER_Neg, PR_Neg, Her2_Neg represent ER negative, PR negative, HER2 negative, respectively. (C) Boxplots showing GS (Y axis) across each module (X axis). (D) Cluster analysis of modules and TNBC to find TNBC-related modules. (E) Scatterplots of correlation between MM (X axis) and GS (Y axis) in cyan module. (F) Scatterplots of correlation between intramodule connectivity (X axis) and GS (Y axis) in cyan module. (G) Scatterplots of correlation between MM (X axis) and GS (Y axis) in yellow module. (H) Scatterplots of correlation between intramodule connectivity (X axis) and GS (Y axis) in yellow module. (I) List of hub genes. WGCNA, weighted gene co-expressed network analysis; MM, module membership.

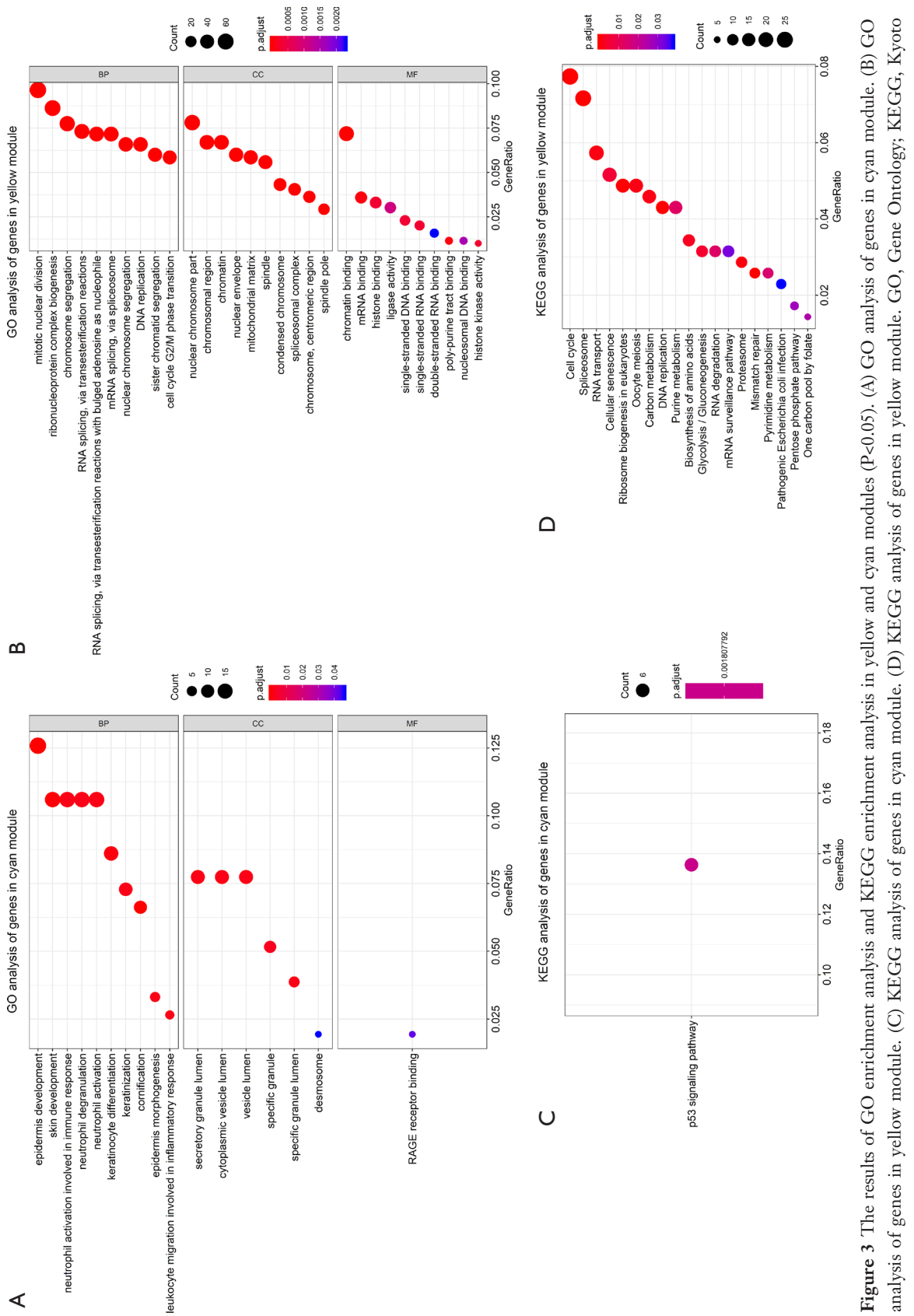

Figure 3 The results of GO enrichment analysis and KEGG enrichment analysis in yellow and cyan modules (P<0.05). (A) GO analysis of genes in cyan module. (B) GO analysis of genes in yellow module. (C) KEGG analysis of genes in cyan module. (D) KEGG analysis of genes in yellow module. GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.

Table 4 Top 10 enriched pathways of DEGs

KEGG term	KEGG ID	P value	Gene symbol
PI3K-Akt signaling pathway	hsa04151	6.51E-06	*EIF4B, RPLR, YWHAB, MAPK3, CDC37, COL4A2, SOS2, CDKN1B, MDM2, GNB4, OSMR, MYC, PPP2R5D, PPP2R5E, CREB3L4, MYB, SPP1*
Metabolic pathways	hsa01100	6.10E-06	*COX6C, DNMT1, SUCLG2, MGAT4A, UXS1, AGL, AKR1B1, UQCRC2, NDUFB1, POLD4, NDUFV2, DCTD, EXT1, DHCR24, PTDSS1, PGM1, CTPS1, GLS, SPTLC2, PNP, ASAH1, ATP5I, CERS2, COX7C, SEPHS1, POLR3K, ATP5G2, B4GALT2, MCCC2, PPCS, UQCRQ, SPR, COQ5*
Protein processing in endoplasmic reticulum	hsa04141	6.10E-06	*PDIA4, HYOU1, STUB1, TRAM1, BAG1, DNAJC1, SAR1B, RRBP1, UBE2J1, LMAN1, CALR, DNAJB11*
Viral carcinogenesis	hsa05203	0.00026	*LTBR, YWHAB, MAPK3, LYN, CDKN1B, MDM2, GTF2H2C_2, SRF, SND1, CREB3L4, ACTN1*
Huntington's disease	hsa05016	0.00076	*LTBR, UQCRQ, MAPK3, LYN, SP1, COX6C, DENND1B, CREB3L4, ATP5G2, UQCRC2*
Oxidative phosphorylation	hsa00190	0.00163	*LTBR, MAPK3, LYN, COX6C, ATP5I, UQCRQ, ATP5G2, UQCRC2*
Parkinson's disease	hsa05012	0.00216	*LTBR, MAPK3, LYN, COX6C, UBE2J1, UQCRQ, ATP5G2, UQCRC3*
Epstein-Barr virus infection	hsa05169	0.00371	*CDKN1B, LYN, MDM2, POLR3K, SHFM1, SND1, MYC, YWHAB, HSPB1*
Proteoglycans in cancer	hsa05205	0.00371	*MAPK3, FLNA, EIF4B, MDM2, SDC4, MYC, RDX, FRS2, SOS2*
Alzheimer's disease	hsa05010	0.00412	*NDUFB1, MAPK3, COX7C, C5orf24, NDUFV2, UQCRQ, DNAJB11, UQCRC2*

DEG, differentially expressed gene; KEGG, Kyoto Encyclopedia of Genes and Genomes.

amegroups.cn/static/public/10.21037atm-20-5989-4.docx). The top BP enrichment GO terms of DEGs were negative regulation of apoptotic process, establishment of protein localization, and protein transport. The top 3 enriched pathways of DEGs were PI3K-Akt signaling pathway, metabolic pathways, and protein processing in endoplasmic reticulum (*Table 4*).

For the cyan module, epidermis development, specific granule lumen, and RAGE receptor binding were most significantly enriched in BP, CC, and MF, respectively (https://cdn.amegroups.cn/static/public/10.21037atm-20-5989-5.docx, *Figure 3A*). Only the p53 signaling pathway was enriched in the cyan module (*Figure 3C*, Table S1). For the yellow module, mitotic nuclear division, chromosomal region, and chromatin binding were most significantly enriched in BP, CC, and MF, respectively (https://cdn.amegroups.cn/static/public/10.21037atm-20-5989-7.docx, *Figure 3B*). Cell cycle, DNA replication, and spliceosome were the top 3 most significantly enriched pathways in the yellow module (https://cdn.amegroups.cn/static/public/10.21037atm-20-5989-8.docx, *Figure 3D*).

Comparing the pathways in DEGs and yellow module, cell cycle, oocyte meiosis, spliceosome, and pathogenic Escherichia coli infection were identified as enriched pathways in both processes.

Prognosis-related genes revealed by survival analysis

According to survival analyses of hub genes in the PPI network and yellow module, we noticed that three key molecules (*HSPB1*, *TPX2*, and *IFI16*) can predict TNBC patients' survival time (*Figure 4*). High expression of *HSPB1* was associated with worse OS in TNBC patients (*Figure 4A*), while low expressions of *TPX2* (*Figure 4B*) and *IFI16* (*Figure 4C*) lead to worse OS in TNBC patients.

Discussion

Application of two different bioinformatics methods including meta-analysis and WGCNA ensured successful identification of novel therapeutic and prognostic biomarkers for TNBC. PPI network analysis showed 11 DEGs with top degrees (*CDH1*, *SP1*, *MYC*, *FAF2*, *IFI16*,

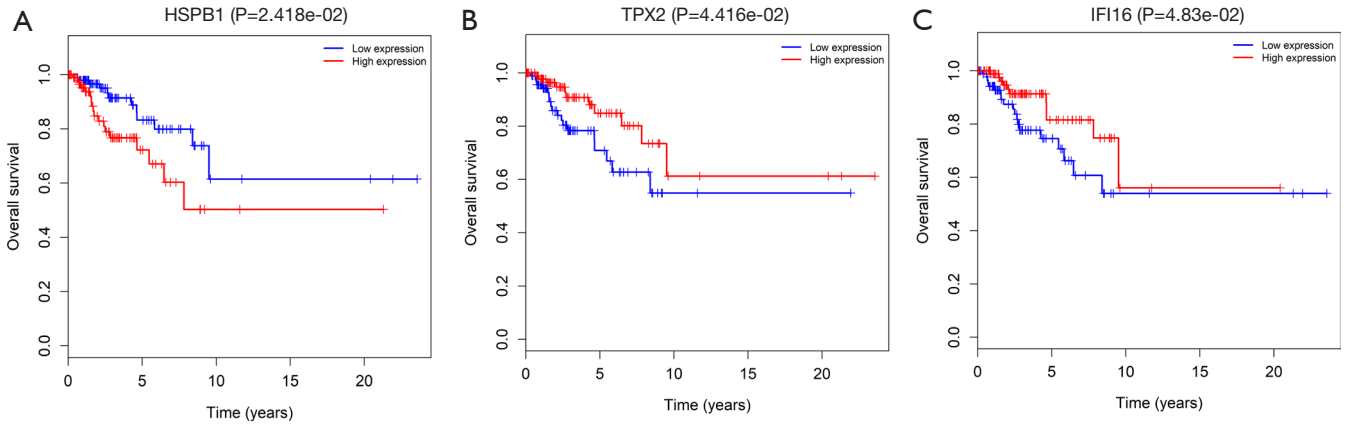

Figure 4 Kaplan-Meier survival curves, which show that expression levels of 3 deregulated genes (A) *HSPB1*, (B) *TPX2* and (C) *IFI16* were associated with prognosis in TNBC patients (unadjusted P value <0.05). Y axis is the overall survival rate and X axis is the survival time. TNBC, triple-negative breast cancer.

MDM2, *AR*, *DBN1*, *HSPB1*, *FLNA*, and *YWHAB*) may play central roles in TNBC. Meanwhile, hub genes (*TPX2*, *CTPS1*, *KIF2C*, *MELK*, and *CDCA8*) in the yellow module of WGCNA may also be crucial for TNBC.

Among these genes, *CDH1*, *MYC*, *AR*, and *MELK* have been reported highly related to TNBC. *CDH1*, as a tumor suppressor gene, encodes E-cadherin protein, which is considered as a promising TNBC marker (18). *CDH1*-promoter demethylation could induce *de novo* E-cadherin expression, suppressing invasion, and metastasis of epithelial tumors (19). Disproportionately increased expression of oncogenic transcription factor MYC was observed in TNBC compared with ER, PR, and HER2 receptor-positive breast cancer (20). Many studies indicated that MYC contributed to cell proliferation and malignant transformation by promoting the expression of cell cycle related genes (21,22). MYC overexpression combining with inactivity of tumor suppressor pathway related to p53 may lead to aberrant tumor growth in TNBC (21). AR, a member of the steroid hormone receptor family, is expressed in the majority of breast carcinoma. The research revealed that tumorigenesis and proliferative change in breast cancer cell lines would occur due to AR dysregulation (23). MELK is a mitotically regulated kinase that could mediate cell survival under metabolic stress. Studies have shown that MELK was overexpressed in basal-like breast cancer and TNBC, which makes MELK a vital target of TNBC (24,25).

In this study, we identified three key genes (*HSPB1*, *TPX2*, and *IFI16*) that were related to the prognosis of TNBC, which had never been reported in previous studies. Contrary to their expression in the meta-analysis, higher expression of *HSPB1* and lower expression of *IFI16* predicted worse OS in TNBC. HSPB1 is a kind of small heat shock protein involved in some cell death pathways like necrosis, apoptosis, or autophagy to protect cells from *lethality*. Overexpression of HSPB1 was associated with increased tumorigenicity, tumor cells metastasis, and *chemotherapeutic* resistance (26). It was reported that high expression of HSPB1 was associated with worse overall survival of breast cancer in general (27), while the relationship between the expression level of HSPB1 and prognosis of breast cancer subtypes is unknown. Our study demonstrated that the upregulation of *HSPB1* was related to the poor prognosis of TNBC. IFI16 is considered as a nuclear pathogen sensor that participates in the innate immune response by sensing foreign DNA (28,29). Our study showed low expression of *TPX2* contributed to worse OS in TNBC. TPX2 is a tubule-associated protein. According to the GO analysis, TPX2 may play an essential role in mitotic nuclear division, cell cycle G2/M phase transition, and other cell cycle-related biological processes.

In order to reveal the underlying molecular mechanisms, pathway enrichment analysis was performed, and the results indicated that cell cycle, oocyte meiosis, spliceosome, and pathogenic Escherichia coli infection had a strong association with TNBC. The first 2 pathways were closely associated with cell growth and death, which were supported by the expression of genes related to proliferation like *MYC*, *IFI16*, *AR*, *MDM2*, *FLNA*, *MELK*, and *CDCA8*.

In conclusion, our integrated analysis of microarray

combined with WGCNA provided promising therapeutic targets for TNBC. Survival analysis uncovered novel prognostic relationships of candidate genes and the overall survival of TNBC patients. Additionally, *HSPB1*, *TPX2*, and *IFI16* have the potential to be used as prognostic biomarkers of TNBC. Our work identified the TNBC-related candidate genes and provided new insights into the underlying mechanisms of TNBC, which is beneficial for improving the outcome of TNBC.

Acknowledgments

Funding: This work was supported by the State Key Laboratory of Chemical Oncogenomics, Technology R & D Funds of Shenzhen, China (Grant No. GJHZ20170314164935502), and Shenzhen Bay Laboratory, Shenzhen, China.

Footnote

Reporting Checklist: The authors have completed the MDAR reporting checklist (available at http://dx.doi.org/10.21037/atm-20-5989).

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/atm-20-5989). The authors have no conflicts of interests to declare.

Ethical Statement: The authors are for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Foulkes WD, Smith IE, Reis-Filho JS, et al. Triple-negative breast cancer. N Engl J Med 2010;363:1938-48.
2. Garrido-Castro AC, Lin NU, Polyak K. Insights into Molecular Classifications of Triple-Negative Breast Cancer: Improving Patient Selection for Treatment. Cancer Discov 2019;9:176-98.
3. Harbeck N, Penault-Llorca F, Cortes J, et al. Breast cancer. Review Nat Rev Dis Primers 2019;5:66.
4. Lee ST, Feng M, Wei Y, et al. Protein tyrosine phosphatase UBASH3B is overexpressed in triple-negative breast cancer and promotes invasion and metastasis. Proc Natl Acad Sci U S A 2013;110:11121-6.
5. Yang L, Wu X, Wang Y, et al. FZD7 has a critical role in cell proliferation in triple negative breast cancer. Oncogene 2011;30:4437-46.
6. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 2005;365:488-92.
7. Ntzani EE, Ioannidis JPA. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. Lancet 2003;362:1439-44.
8. Ramasamy A, Mondry A, Holmes CC, et al. Key Issues in Conducting a Meta-Analysis of Gene Expression Microarray Datasets. PLoS Med 2008;5:e184.
9. Yin L, Cai Z, Zhu B, et al. Identification of Key Pathways and Genes in the Dynamic Progression of HCC Based on WGCNA. Genes (Basel) 2018;9:92.
10. Carvalho B, Bengtsson H, Speed TP, et al. Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. Biostatistics 2007;8:485-99.
11. Kang DD, Sibille E, Kaminski N, et al. MetaQC: objective quality control and inclusion/exclusion criteria for genomic meta-analysis. Nucleic Acids Res 2012;40:e15.
12. Wang X, Kang DD, Shen K, et al. An R package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection. Bioinformatics 2012;28:2534-6.
13. Stark C, Breitkreutz BJ, Reguly T, et al. BioGRID: a general repository for interaction datasets. Nucleic Acids Res 2006;34:D535-9.
14. Keshava Prasad TS, Goel R, Kandasamy K, et al. Human Protein Reference Database--2009 update. Nucleic Acids Res 2009;37:D767-72.
15. Kohl M, Wiese S, Warscheid B. Cytoscape: Software for Visualization and Analysis of Biological Networks. In: Hamacher M, Eisenacher M, Stephan C, editors. Data Mining in Proteomics: From Standards to Applications. Totowa, NJ: Humana Press, 2011:291-303.
16. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC

Bioinformatics 2008;9:559.
17. Yu G, Wang LG, Han Y, et al. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16:284-7.
18. Ricciardi GRR, Adamo B, Ieni A, et al. Androgen Receptor (AR), E-Cadherin, and Ki-67 as Emerging Targets and Novel Prognostic Markers in Triple-Negative Breast Cancer (TNBC) Patients. PLoS One 2015;10:e0128368.
19. Lopes N, Carvalho J, Durães C, et al. 1Alpha,25-dihydroxyvitamin D3 induces de novo E-cadherin expression in triple-negative breast cancer cells by CDH1-promoter demethylation. Anticancer Res 2012;32:249-57.
20. Carey JPW, Karakas C, Bui T, et al. Synthetic Lethality of PARP Inhibitors in Combination with MYC Blockade Is Independent of BRCA Status in Triple-Negative Breast Cancer. Cancer Res 2018;78:742-57.
21. Fallah Y, Brundage J, Allegakoen P, et al. MYC-Driven Pathways in Breast Cancer Subtypes. Biomolecules 2017;7:53.
22. Carey JPW, Keyomarsi K. Leveraging MYC as a therapeutic treatment option for TNBC. Oncoscience 2018;5:137-9.
23. Mina A, Yoder R, Sharma P. Targeting the androgen receptor in triple-negative breast cancer: current perspectives. Onco Targets Ther 2017;10:4675-85.
24. Edupuganti R, Taliaferro JM, Wang Q, et al. Discovery of a potent inhibitor of MELK that inhibits expression of the anti-apoptotic protein Mcl-1 and TNBC cell growth. Bioorg Med Chem 2017;25:2609-16.
25. Pitner MK, Taliaferro JM, Dalby KN, et al. MELK: a potential novel therapeutic target for TNBC and other aggressive malignancies. Expert Opin Ther Targets 2017;21:849-59.
26. Acunzo J, Katsogiannou M, Rocchi P. Small heat shock proteins HSP27 (HspB1), αB-crystallin (HspB5) and HSP22 (HspB8) as regulators of cell death. Int J Biochem Cell Biol 2012;44:1622-31.
27. Choi SK, Kam H, Kim KY, et al. Targeting Heat Shock Protein 27 in Cancer: A Druggable Target for Cancer Treatment? Cancers (Basel) 2019;11:1195.
28. Jakobsen MR, Bak RO, Andersen A, et al. IFI16 senses DNA forms of the lentiviral replication cycle and controls HIV-1 replication. Proc Natl Acad Sci U S A 2013;110:E4571-80.
29. Kerur N, Veettil Mohanan V, Sharma-Walia N, et al. IFI16 acts as a nuclear pathogen sensor to induce the inflammasome in response to Kaposi Sarcoma-associated herpesvirus infection. Cell Host Microbe 2011;9:363-75.

Cite this article as: Cao W, Jiang Y, Ji X, Guan X, Lin Q, Ma L. Identification of novel prognostic genes of triple-negative breast cancer using meta-analysis and weighted gene co-expressed network analysis. Ann Transl Med 2021;9(3):205. doi: 10.21037/atm-20-5989

Supplementary

Table S1 KEGG analysis of genes in cyan module

ID	Description	GeneRatio	BgRatio	P value	p.adjust	qvalue	Gene ID	Count
hsa04115	p53 signaling pathway	6/44	72/4789	4.20E-05	0.001808	0.001725921	*CD82/SERPINB5/PERP/STEAP3/BID/CDKN2A*	6

http://dx.doi.org/10.21037/atm-20-5989