Identification of Potential Diagnostic Genes for Bladder Cancer by Bioinformatic Analysis

Yongxing Peng
yingtan's people hospital
Minqin Mao
JiangXi teachers college
Zhonglai Li
Yingtan people's hospital
Qipeng Xia
Yingtan people's hospital
Honghua Tong (2002022006@jxsfgz.com)
YingtanPeople's Hospital https://orcid.org/0000-0002-5535-4063

Primary research

Keywords: bladder cancer, biomarkers, bioinformatics, area under the curve

Posted Date: November 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-978784/v1

Abstract

Background: Cytology and transurethral cystoscopy constitute the gold standard for the diagnosis of bladder cancer (BC). However, some minor lesions cannot be detected in time with these techniques, resulting in a high rate of missed diagnosis. Finding biomarkers that are economical, convenient, sensitive, and specific has become an urgent priority.

Methods: Gene expression profile data from BC and normal bladder tissue were downloaded from the Gene Expression Omnibus (GEO) database and used as a training set to screen for differentially expressed genes (DEGs). The bladder gene expression and related clinical data derived from The Cancer Genome Atlas (TCGA) and Genotype-Tissue Expression (GTEx) databases were used as a validation set. The effectiveness of the DEGs as diagnostic criteria was verified in terms of gene expression, gene mutation and diagnostic efficiency.

Results: Two upregulated and eight downregulated hub genes were identified by screening. In terms of gene expression, the expression levels of these genes were significantly different between bladder cancer tissues and normal tissues. In terms of clinical diagnostic efficacy, TOP2A had the highest single diagnostic value, while the combinations of TOP2A/CNN1, TOP2A/ISG15/CNN1 and TOP2AISG15/ACTG2 had the largest area under the curve (AUC) among two- or three-indicator combinations.

Conclusion: TOP2A, either alone or as part of a combination, has notable diagnostic advantages. However, this still needs to be confirmed in a larger sample with further biological experiments.

1. Introduction

Worldwide, bladder cancer (BC) accounts for 3% of all incident cases of cancer. Approximately 150 thousand patients die from BC each year [1–2]. Although cytology and transurethral cystoscopy constitute the gold standard for diagnosing BC, BC is easily missed due to tumour heterogeneity and some subtle pathological changes such as small papillary tumours and micrometastasis [3]. Therefore, it has been observed that approximately 1/4 of newly diagnosed BC cases in clinical practice involve muscle infiltration [4–5]. In addition, non-muscle-invasive BC has a high risk of recurrence after treatment, at more than 50%. More seriously, 30% of cases of noninvasive bladder cancer transform into invasive bladder cancer [6]. Because of the high risks of recurrence and progression due to the uncertainty of diagnosis, BC leads to repeated traumatic examination and treatment, resulting in a heavy financial burden for patients [2, 7].

In light of this situation, molecular markers, which can safely and effectively diagnose BC without trauma, have become another ideal choice. However, to date, few widely accepted molecular markers have been applied for clinical diagnosis. For example, nuclear mitotic apparatus protein (NMP22), which has been approved as a BC biomarker by the US Food and Drug Administration (FDA), still has the problem of relatively low sensitivity and even needs to be combined with cystoscopy [8–10]. Conversely,

a screening analysis of 1502 people with a high risk of BC showed that only 85 subjects were positive for NMP22. A further examination of 69 suspected BC patients who were willing to be tested revealed that only 3 patients were diagnosed with BC [11]. Just as importantly, NMP22 is also positive in cystitis, urinary calculi and other pathological conditions [12]. This highlights the desperate need to identify novel reliable molecular markers.

With the great advancement of high-throughput technologies and the establishment of various public medical databases such as the Cancer Genome Atlas (TCGA https://cancergenome.nih.gov) and Gene Expression Omnibus (GEO https://www.ncbi.nlm.nih.gov) online databases, tissue differentially expressed gene (DEG) analysis and functional enrichment analysis have been applied to many human cancers. In the current study, the gene expression profile data of BC and normal bladder tissue were extracted from the GEO database as a training set, and the DEGs between the tissues were taken as the research objects. Expression profiles and corresponding clinical data related to bladder tissue were obtained from the databases of TCGA and Genotype-Tissue Expression (GTEx, https://gtexportal.org) and used as a validation set. Using the Human Protein Atlas (HPA https://www.proteinatlas.org) and other databases to systematically study the efficiency of screening genes in diagnosis, we identified potential biomarkers for the diagnosis of BC.

2. Materials And Methods

2.1 Microarray data

In this study, three sets of human genetic screening data were downloaded from the GEO database (www.ncbi.nlm.nih.gov/geo/); these datasets were based on the PL96 (Affymetrix Human Genome U133A Array), PL570 (Affymetrix Human Genome U133 Plus 2.0 Array) and GSE13507 datasets, respectively. Table 1 details the specifics of each of the three datasets, including the numbers of normal and BC specimens and the year of the trial. For different GEO data on the same platform, we followed a standard operating procedure and normalization method based on other scholars' studies [13–14]. Meanwhile, 407 BC and 28 normal samples collected from TCGA and GTEx were used as a validation set [15]. In the screening process, we removed the probes corresponding to multiple molecules. In this case, only the probe with the largest signal value was retained.

Table 1
Details for GEO bladder cancer datasets

Group	GEO ID	GPL ID	Normal	Tumor	Year	Reference
	GSE17906	GPL570	2	1	2009	Pascal et al.
	GSE30522	GPL570	1	1	2011	Liu et al.
	GSE7476	GPL570	3	9	2007	Mengual et al.
Total GPL570 Set			6	11		
	GSE3167	GPL96	9	51	2005	Dyrskjøt et al.
	GSE5287	GPL96	0	30	2006	Als et al.
	GSE37317	GPL96	0	19	2012	Smith et al.
Total GPL96 Set			9	100		
GSE13507	GSE13507	GPL6102	68	188	2008	Kim et al.

2.2 Screening of differentially expressed genes (DEGs)

After the normal control group and the BC group were compared and analysed using the limma package in R (v.3.5.3), DEGs were obtained, and the results were deduplicated. Screening criteria were as follows: absolute value of $\log_2$ gene expression fold change (FC) ≥ 1, adjP < 0.05. Volcano plots were drawn with the "ggplot2" R package. To eliminate the influence of mixed false-positive expression genes and screen the prediction biomarkers for subsequent clinical diagnosis, only common differentially expressed genes were selected for analysis. In the process, the "VennDiagram" R package was used to draw Venn diagrams to achieve this purpose.

2.3 PPI network construction and module analysis

The online database STRING (https://string-db.org/) was used to analyse the protein interaction network of common DEGs (minimum required interaction score=0.40), and the results were imported into Cytoscape software for visualization and correlation analysis. The hub genes were screened by using Cytohubba, a plug-in of Cytoscape software.

2.4 Gene Ontology and pathway enrichment analyses

The online Gene Ontology (GO) database (http://geneontology.org/) is an effective bioinformatics tool that provides a systematic understanding of the biological functions of gene products at the human cell level. KEGG (https://www.kegg.jp/) is a database that uses genetic information to calculate and speculate on higher-level and more complex cell activity as well as biological behaviour. Before analysis, wide genome annotation for the human R package "org.Hs.eg.db" was used to transform the gene symbol codes for all data into Entrez IDs. KEGG pathway analysis and Gene Ontology (GO) analysis were

employed using the clusterProfiler package to better understand the intrinsic biomedical functions and essential characteristics of genes.

2.5 Oncomine database analysis

Oncomine (https://www.oncomine.org/resource/login.html) is the largest oncogene chip database and integrated data mining platform in the world [16], which aims to mine the key genes of various cancers. Using the public datasets available, DEG expression in BC tissues and noncancerous tissues was compared. The P value was generated using Student's t-test (p value ≤ 0.01, fold change 1).

2.6 GEPIA database analysis

GEPIA (Gene Expression Profiling Interactive Analysis) is a web server that uses integrated and standardized TCGA and GTEx gene expression information to identify the differences and characteristics between cancer and normal gene expression [17]. In the present study, screening gene expression values of human BC tissue originating from TCGA normal and GTEx databases were matched with those in normal tissues, and we verified the gene expression difference among different tissues through the online GEPIA platform.

2.7 Gene expression verification with R software

This stage includes three parts: (1) The TCGA expression difference in paired samples (cancer and adjacent tissues) from BC patients; (2) The expression discrepancy among cancer samples with different pathological stages and control normal samples; (3) Expression differences among cancer samples with different histological grades and normal control samples. To achieve the above objectives, validation data were downloaded from level 3. HTSeq-TPM RNAseq in the TCGA BLCA (bladder urothelial carcinoma) project. After log2 conversion, it was calculated and visualized using paired t-test or Kruskal–Wallis test and ggplot2 package of R statistical software.

2.8 Genetic alteration and HPA database analysis

The alteration frequency and mutation type of screening genes in BC were compiled using cBioPortal (https://www.cbioportal.org/). By inputting the target genes in the "quick select" option, details and summaries of mutation types and gene amplification in BC can be observed in the "Cancer Type Summary" and "OncoPrint" modules, respectively. In addition, the "Comparison/Survival" module was used to obtain the difference in overall survival between the altered and unaltered observed gene groups, and a Kaplan–Meier diagram with the log-rank test P value was drawn. To verify the protein expression of the studied genes, biopsy immunohistochemical micrographs of tumour or normal tissues were acquired from the Human Protein Atlas (HPA http://www. proteinatlas.org/).

2.9 Propensity score matching (PSM) and random forest analysis

The specimens collected from the above TCGA and GTEX databases were divided into a normal group and a BC group. The gender and age in the data were taken as the variables to be matched, and the genes

to be selected were taken as the screening indicators for cancer diagnosis. Since the age in the GTEX database is presented in 10-year bins, we divided all data into five groups (patients under the age of 40 years were in the first group, and those over the age of 70 years were in the fifth group). In addition, since the number of people in the cancer group was much larger than that in the normal group, the two groups were matched by propensity score (PSM) at a ratio of 1:5 according to the principle of statistical proportion selection. After the matchit, tableone and random forest packages were downloaded [18], random forest analysis was carried out in R software.

2.10 Diagnostic ROC

The aforementioned samples collected from TCGA and GTEx databases were divided into a normal group and a BC group. First, ROC analysis of a single gene was carried out. After $\log_2$ transformation of TPM RNA-seq expression data of the target gene in each sample, the transformed values were included in each group for calculation. The area under the curve (AUC) and visualization of each gene predicting cancer were calculated by using the pROC (version 1.17.0.1) and ggplot2 R software packages, respectively. Then, a logistic regression model was used for all genes. Since the screening genes had been determined at this step, the formula is as follows:

glm (formula = status ~ TOP2A + ISG15 + CNN1 + MYH11 + ACTA2 + MYLK + LMOD1 + ACTC1 + ACTG2 + TAGLN, data= data, family = binomial).

A model was established according to the regression coefficient of each gene, representing the value generated by all indicators through the model, and the value was used to predict a final outcome, which was the joint indicator ROC analysis. Finally, two or three random genes were combined for ROC analysis. Under the working conditions of R software, merged polygenic data can be modelled by the glm function, joint ROC analysis by the pROC package, and visualization by the ggplot2 package. The prediction ability of genes was identified by the numerical value of AUC.

2.11 GSEA and gene correlation analysis

GSEA (gene set enrichment analysis) is an analytical method that derives gene sets to ascertain diverse biological functions between two phenotypes [19–20]. After loading the clusterprofiler (version 3.14.3), stat (version 3.14.3) and ggplot2 packages under the working state of R software, GSEA, Pearson correlation analysis and visualization operation were executed for the gene screened by random forest analysis and joint diagnosis ROC curve analysis. It is generally accepted that the conditions of false discovery rate (FDR) < 0.25 and i < 0.05 indicate significant enrichment.

Table 2 The assignment of each gene after logistic regression of all genes was included.

Variable	Coefficient	Standard error SE	P value
Intercept	9.028	8.672	0.298
TOP2A	-0.872	0.439	0.047 *
ISG15	-1.356	0.630	0.031 *
CNN1	1.820	2.100	0.386
MYH11	0.491	1.708	0.774
ACTA2	4.397	2.427	0.070
MYLK	-2.446	1.030	0.018 *
LMOD1	1.290	1.907	0.499
ACTC1	-0.722	0.382	0.059
ACTG2	2.469	1.850	0.182
TAGLN	-6.740	2.812	0.017 *

3. Results

3.1 Identification of DEGs

To exclude the batch effect, the gene expression levels of multiple GSE sets in the same PL570 or PL96 platform were standardized. The differences before and after standardization are shown in Fig. 1A and 1B. After limma calculation and inclusion condition screening, 271 upregulated genes and 42 downregulated genes were detected in the PL96 platform dataset. In addition, 172, 66 upregulated and 680,392 downregulated DEGs were selected from the PL570 and GSE13507 datasets, respectively. Then, volcano plots using Prism software (Fig. 1C) showed the meaningful DEGs screened from the three target datasets, and Venn diagrams (Fig. 1D) presented the 2 upregulated and 16 downregulated overlapping genes, which were calculated by the VennDiagram software package.

3.2 PPI network analysis

A network diagram was established for 16 downregulated genes through the PPI online network (Fig. 2A). These genes were intersected using Cytohubba in Cytoscape to identify hub genes. To ensure the rationality of the study, 12 methods, including the MCC method, were used to calculate the top 10 ranked genes. After different methods of summary and analysis, 8 downregulated genes were included for further research **(Additional file1)**. In the TCGA database, Spearman correlation analysis of 8 downregulated hub genes and 2 upregulated genes (Fig. 2B) was performed using the R software pheatmap package. The downregulated genes were closely related, but the correlation with the

upregulated genes was not obvious. Figure 2C showed the log2 expression FC values of the screened hub genes in different data sets

3.3 Gene Ontology and pathway enrichment analyses

With *Homo sapiens* as the background, the clusterProfiler software package analysed the screened genes and assessed their enrichment patterns. Next, BP, CC and MF enrichment analysis was performed on the downregulated DEGs (Fig. 2D). The downregulated genes were mainly enriched in muscle contraction and muscle system processes (ontology: BP), contractile fibres (ontology:CC) and structural constituents of muscle (ontology: MF). KEGG pathway analysis showed that these integrated DEGs were mainly enriched in vascular smooth muscle contractions.

3.4 Gene expression analysis

First, the Oncomine database was searched to compare the mRNA transcription levels of target genes in BC samples (Fig. 3A). The results showed that the expression of TOP2A and ISG15 increased in cancer samples and decreased in control samples, while the expression of CNN1 and other genes showed the opposite outcome. In Sanchez-Carbayo's study [21], TOP2A expression in BC was significantly increased and was 9.402 times higher than that in normal tissues (Fig. 3B). Another typical study [22] also compared the expression of CNN1 among normal (58 cases) and benign tissues (10 cases), superior BC (126 cases) and infiltrating bladder urothelial carcinoma (62 cases) and found that its expression level decreased in turn (Fig. 3C). The expression results calculated based on the GEPIA database are shown in Figue 3D. TOP2A and ISG15 were highly expressed in cancer tissues, and other genes were poorly expressed. This was exactly the same as the result found by using the Oncomine database. Interestingly, the following paired t-test results of gene expression in adjacent and cancer tissues by R software (Fig. 3E) were still completely consistent with the first two studies, and the P values were all less than 0.001. To further test our hypothesis, we compared the expression of screening genes in cancer samples with different pathological stages and histological grades with that in normal controls. The results showed that although TOP2A was not overexpressed in low-grade human bladder cancer versus disease-free tissue, there were significant statistical expression divergences of all genes between normal tissue and different pathological stages (Fig. 3F) or histological grades of tumour tissue (Fig. 3G).

3.5 Genetic alteration and HPA database analysis

As shown in Figue 4A, the mutation rate of the target genes was between 1% and 8%. Among them, the missense mutation rate of NYH11 was the highest, which was more obvious than other mutations, such as deletion mutations. From the total mutation rate of related genes, the mutation rate of the Cornell & Trento study in 72 patients was 43.06%. Likewise, in the TCGA database containing 412 patients, the proportion still reached 27.91% (Fig. 4B). Within the "Comparison/Survival" module, the impact of ISG15 and all downregulated genes on health and survival can be explained by gene variation (Fig. 4C-D). At the protein level, immunohistochemical staining of the screened genes in the HPA database showed that the expression of TOP2A and ISG15 increased in BC, while the expression of MYH11 and ACTA2 decreased

(Fig. 5). In addition, the expression of expected downregulated genes, such as CNN1, was not obvious in either cancer tissues or normal tissues; therefore, it was extremely difficult to compare them.

3.6 PSM and random forest analysis

After matching, Figue 6A shows that the average difference in propensity scores for gender and age decreased to 0. All 28 original normal samples were included, while only 122 cancer tissue samples remained. In age Group 1, due to a shortage of samples, only 22 cases were included in the analysis instead of the intended 40. The similarity of the two matched groups was identified by Figure 6B, and it could be found that they have good similarity. The matched data were then subjected to random forest analysis. The results indicated that the mean decrease in Gini (MDG) results of TOP2A were the largest, followed by MYH11 and MOD1 (Fig. 6C). It is generally accepted that the MDG is directly proportional to the prediction ability of the screened gene. Therefore, it could be considered that TOP2A has greater value in the diagnosis of BC than other genes. Fig. 6D came from the multidimensional scale diagram of R software. The red and blue points represent the cases in the cancer group and control group, respectively. The more obvious the aggregation of points of the same colour, and the greater the distance between different colour points in the diagram, the more significant the calculation result is.

3.7 Diagnostic ROC

The results of independent diagnostic ROC curve analysis showed that each gene had valuable predictive ability (Fig. 7A). Logistic regression models of all genes showed that TOP2A, ISG15, MYLK and TAGLN had statistical diagnostic value (Table 2). The AUC calculated by from the diagnostic ROC of all genes combined was 0.996 (Fig. 7B). In view of clinical practicality, we tested combinations of 2 or 3 genes for joint diagnosis and found that the combinations of TOP2A/CNN1, TOP2A/ISG15/ACTG2 and TOP2A/ISG15/CNN1 had the largest area under the ROC curve (Fig. 7C-D).

3.8 GSEA and gene correlation analysis

Because TOP2A holds a special position in the diagnosis of BC, TOP2A was specifically studied. GSEA showed that the REACTOME_SIGNALING_BY_RHO_GTPASES and REACTOME_M_PHASE gene sets were significantly enriched (Fig. 7E). A gene correlation study (Fig. 7F) showed that 10 genes, FABP7 and SOX14, were most significantly correlated, of which 8 genes were statistically significant.

4. Discussion

Because BC is prone to relapse, easily progresses and has high treatment costs, it is particularly imperative to identify cancer biomarkers that can reflect the biological functions of BC. Konety et al. posited that the ideal bladder cancer marker would not only be safe, economical and stable but also have low false-positive and false-negative rates [23]. Meanwhile, diagnostic biomarkers can take many forms

and may be lipids, proteins or DNA in this study, and their expression levels are highly correlated with the risk of bladder cancer [24].

In the current research, the datasets of the same PL570 and PL96 platforms were merged and standardized, and then the limma package was used to find DEGs on these two platforms combined with the GSE13507 dataset. After the intersection of the DEGs of the three datasets, two common upregulated genes (TOP2A and ISG15) and 11 downregulated DEGs were found. Thereafter, GO and KEGG enrichment analyses of downregulated DEGs were carried out, revealing that the DEGs were associated with a decline in muscle system contraction. Next, eight downregulated genes identified through the PPI network, together with the above two upregulated genes, were selected as hub genes. To verify the diagnostic validity, the transcriptional level, mutation rate and protein expression of hub genes in BC tissues were compared with those of normal tissues in TCGA and GTEx datasets, and satisfactory results were found. To validate their clinical applicability, the hub genes were analysed by random forest analysis and combined ROC analysis, and TOP2A was found to have conspicuous diagnostic advantages. In random forest analysis, TOP2A had the highest MDG value, which meant that its separate diagnostic value was the most significant. Among combinations of two indexes, TOP2A and CNN1 were the highest-performing pair. Among three-gene combinations, TOP2A combined with CNN1 plus ISG15 or ISG15 plus ACTG2 had the largest AUC.

TOP2A, also known as topoisomerase II α, is an enzyme that changes the topological state of DNA by breaking double-stranded DNA during mitosis and participates in the process of DNA transcription and replication in cell proliferation [25–26]. Generally, cancer is a disease characterized by malignant proliferation and death regulation disorder of tumour cells. Therefore, abnormal TOP2A overexpression is closely related to cancer. Previously, its overexpression was reported in ovarian cancer, gastric cancer, colorectal cancer and other cancer tissues [27–29]. In the present study, the expression of TOP2A in BC and normal tissues was compared in the GEPIA, Oncomine, HPA and R software platforms, and it was found that the expression of TOP2A in distinct stages of bladder cancer was significantly different from that in normal tissues. GSEA also showed mitotic phase and Rho GTPase enrichment. These results were consistent with those of Gao et al [30–31]. Moreover, similar findings were found in mouse bladder cancer models [32]. On this basis, Zeng [33] further demonstrated that the invasiveness, proliferation and migration ability of TOP2A knockout BC cells were markedly restrained by xenograft tumour formation in nude mice and other tests. Regarding the clinical application of TOP2A for BC, Kim's report is the only one to date regarding urinary cell-free DNA [34]. His research showed that the expression level of TOP2A in BC patients was higher than that in normal persons and haematuria patients. In addition, the areas under the ROC curves of TOP2A for BC, non-muscle-invasive BC and muscle-invasive BC were 0.741, 0.701 and 0.838, respectively.

Although few studies have been performed on BC, CNN1 was found to be expressed at low levels in an array of cancer tissues [35–37]. Menéndez [38] reported that CNN1 can inhibit tumour formation by mediating a variety of angiogenic factors. Scratch and wound healing tests also showed that the invasion and migration ability of CNN1-overexpressing lung squamous cell carcinoma cells decreased

significantly [39]. ISG15 overexpression has been reported to promote distant metastasis of mouse liver cells. However, in fact, its effects on different tumour cells are contradictory [40–42] and still need to be further clarified by scholars.

Finally, it is worth pointing out that our research may have some limitations. First, DEGs were detected in resected tissues using microarray or sequencing technology. Unfortunately, it is unreasonable for these two techniques to be widely used in clinical practice. The search must continue for an ideal technology that can find genetic differences in blood or urine samples economically and quickly. Another intrinsic limitation is that the specific cancer-promoting mechanisms of the screening biomarkers are still obscure; thus, further in vivo and in vitro experiments are necessary. Last but not least, there were relatively few clinical diagnostic tests for screening gene biomarkers. In Kim's study on the diagnostic function of TOP2A in bladder cancer, the sample size of 83 cancer patients still proved to be insufficient. External validation in diverse and much larger-scale populations with longer follow-up periods may be beneficial in further studies.

In summary, the purpose of the present study was to identify the hub genes involved in the development of BC through a comprehensive bioinformatics analysis. The diagnostic effectiveness of the screened genes was verified by expression, mutation rate and clinical application. Our findings indicated that TOP2A had prominent diagnostic value for BC. Nevertheless, larger populations and biological trials are needed to verify this finding.

Abbreviations

BC
Bladder cancer
GEO
Gene Expression Omnibus
DEG
Differentially expressed genes
TCGA
The Cancer Genome Atlas
GTEx
Genotype-Tissue Expression
AUC
Area under the curve
FDA
Food and Drug Administration
HPA
Hunman Protein Atlas
NMP22
Nuclear mitotic apparatus protein. FC:Fold change

GO
Gene ontology
KEGG
Kyoto encyclopaedia of genes and genomes
Gepia
Gene expression profiling interactive analysis
BLCA
Bladder urothelial carcinoma
PSM
Propensity score matche
GSEA
Gene set enrichment analysis
MDG
Mean decrease Gini.

Declarations

Acknowledgements

Not applicable.

Authors' contributions

HT and YP conceived and designed the experiments. YP and MM acquired the data. ZL and QX analysed data. YP draft this manuscript. All authors read and approved the final manuscript.

Funding

No funds, grants, or other support was received.

Availability of data and materials

The datasets analyzed during the current study are available in public databases such as TCGA, geo, GTEX, etc, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

Ethics approval and consent to participate

Not applicable.

Consent for publication

All the authors have read and approved the paper and declare no potential conficts of interest in the paper. All the authors agree to publish this paper.

Competing interests

The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394-424.
2. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ. Cancer statistics, 2009. CA Cancer J Clin. 2009;59(4):225-49.
3. Soloway MS, Sofer M, Vaidya A. Contemporary management of stage T1 transitional cell carcinoma of the bladder. J Urol. 2002;167(4):1573-83.
4. Rhijn BWG, Poel HG, Kwast TH. Urine markers for bladder cancer surveillance: a systematic review. Eur Urol. 2005;47(6):736-48.
5. Witjes JA, Lebret T, Comperat EM, Cowan NC, Santis MD, Bruins HM, et al. Updated 2016 EAU guidelines on muscle-invasive and metastatic bladder Cancer. Eur Urol. 2017; 71(3):462-75.
6. Saad A, Hanbury DC, McNicholas TA, Boustead GB, Morgan S, Woodman AC. A study comparing various noninvasive methods of detecting bladder cancer in urine. BJU Int. 2002;89(4):369-73.
7. Yabroff KR, Lamont EB, Mariotto A, Warren JL, Topor M, Meekins A, et al. Cost of care for elderly cancer patients in the United States. J Natl Cancer Inst. 2008;100(9):630-41.
8. Grossman HB, Messing E, Soloway M, Tomera K, Katz G, Berger Y, et al. Detection of bladder cancer using a point-of-care proteomic assay. JAMA.2005;293(7):810-6.
9. Ponsky LE, Sharma S, Pandrangi L, Kedia S, Nelson D, Agarwal A, et al. Screening and monitoring for bladder cancer: refining the use of NMP22. J Urol.2001; 166(1):75-8.
10. Starke N, Singla N, Haddad A, Lotan Y. Long-term outcomes in a high-risk bladder cancer screening cohort. BJU Int. 2016;117(4):611-17.
11. Lotan Y, Elias K, Svatek RS, Bagrodia A, Nuss G, Moran B, et al. Bladder cancer screening in a high risk asymptomatic population using a point of care urine based protein tumor marker. J Urol. 2009;182(1):52–7.
12. Lotan Y, O'Sullivan P, Raman JD, Shariat SF, Kavalieris L, Frampton C, et al. Clinical comparison of noninvasive urine tests for ruling out recurrent urothelial carcinoma. Urol Oncol. 2017;35(8):531.e15-.e22.
13. Sun TQ, Guan Q, Wang YJ, Qian K, Sun WY, Ji QH, et al. Identification of differentially expressed genes and signaling pathways in papillary thyroid cancer: a study based on integrated microarray and bioinformatics analysis. Gland Surg. 2021; 10(2):629-44.

14. Gu Z, Eils r, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016; 32(18): 2847-49.
15. Vivian J, Rao AA, Nothaft FA, Ketchum C, Armstrong J, Novak A, et al. Toil enables reproducible, open source, big biomedical data analyses. Nat Biotechnol. 2017;35(4):314-6.
16. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu JJ, Briggs BB, et al. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia. 2007;9(2):166-80.
17. Tang ZF, Li CW, Kang BX, Gao G, Li C, Zhang ZM. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Res. 2017;45(W1):W98-102.
18. Hua L, Li DG, Lin H, Li L, Li X, liu ZC. The correlation of gene expression and co-regulated gene patterns in characteristic KEGG pathways. J Theor Biol. 2010;266(2): 242-9.
19. Yu GC, Wang LG, Han YY, H QY, Liu ZC. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics. 2012; 16(5): 284-7.
20. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005; 102(43): 15545-50.
21. Sanchez-Carbayo M, Socci ND, Lozano J, Saint F, Cordon-Cardo C. Defining molecular profiles of poor outcome in patients with invasive bladder cancer using oligonucleotide microarrays. J Clin Oncol. 2006;24(5):778-89.
22. Lee JS, Leem SH, Lee SY, Kim SC, Park ES, Kim SB, et al. Expression signature of E2F1 and its associated genes predict superficial to invasive progression of bladder tumors. J Clin Oncol. 2010;28(16):2660-7.
23. Konety BR. Molecular markers in bladder cancer: A critical appraisal. Urol Oncol. 2006;24(4):326-37.
24. Leiblich A. Recent Developments in the Search for Urinary Biomarkers in Bladder Cancer. Curr Urol Rep. 2017;18(12): 100.
25. Jain M, Zhang L, He M, Zhang YQ, Shen M, Kebebew E. TOP2A is overexpressed and is a therapeutic target for adrenocortical carcinoma. Endocr Relat Cancer. 2013;20(3):361–70.
26. Strausfeld U, Richter A. Simultaneous purification of DNA topoisomerase I and II from eukaryotic cells. Prep Biochem. 1989;19(1):37-48.
27. Faggad A, Darb-Esfahani S, Wirtz R, Sinn B, Sehouli J, Könsgen D, et al. Topoi-somerase IIalpha mRNA and protein expression in ovarian carcinoma: correlation with clinicopathological factors and prognosis. Mod Pathol. 2009; 22: 579-88.
28. Coss A, Tosetto M, Fox EJ, Sapetto-Rebow B, Gorman S, Kennedy BN, et al. Increased topoisomerase IIalpha expression in colorectal cancer is associated with advanced disease and chemothera-peutic resistance via inhibition of apoptosis. Cancer Lett. 2009; 276(2): 228-38.
29. Liang Z , Zeng X , Gao J , Wu S , Wang P , Shi XH, et al. Analysis of EGFR, HER2, and TOP2A gene status and chromosomal polysomy in gastric adenocarcinoma from Chinese patients. BMC Cancer.

2008; 8: 363.
30. Gao X, Chen YY, Chen M, Wang SL , Wen XH, Zhang SF. Identification of key candidate genes and biological pathways in bladder cancer. PeerJ. 2018 ;6:e6036.
31. Li S, Liu XP , Liu TZ, Meng XY, Yin XH, Fang C, et al. Identification of Biomarkers Correlated with the TNM Staging and Overall Survival of Patients with Bladder Cancer. Front Physiol. 2017;8:947.
32. Lu Y, Liu P, Wen W, Grubbs CJ, Townsend RR, Malone JP, et al. Cross-species comparison of orthologous gene expression in human bladder cancer and carcinogen-induced rodent models. Am J Transl Res. 2010 ;20:3(1):8-27.
33. Zeng SX, Liu AW, Dai LH, Yu XW, Zhang ZS, Xiong Q, et al. Prognostic value of TOP2A in bladder urothelial carcinoma and potential molecular mechanisms. BMC Cancer. 2019;19(1):604.
34. Kim YH, Yan CN, Lee IS, Piao XM, Byun YJ, Jeong P, et al. Value of urinary topoisomerase-IIA cell-free DNA for diagnosis of bladder cancer. Investig Clin Urol. 2016;57(2):106-12.
35. Sasaki Y , Yamamura H, Kawakami Y, Yamada T, Hiratsuka M, Kameyama M, et al. Expression of smooth muscle calponin in tumor vessels of human hepatocellular carcinoma and its possible association with prognosis. Cancer. 2002;94(6): 1777-86.
36. Wang Z, Li TE, Chen M, Pan JJ, Shen KW. miR-106b-5p contributes to the lung metastasis of breast cancer via targeting CNN1 and regulating Rho/ROCK1 pathway. Aging (Albany NY). 2020 12(2):1867-87.
37. Mulas JM, Reymundo C, Monteros AE , Millán Y, Ordás J. Calponin expression and myoepithelial cell differentiation in canine, feline and human mammary simple carcinomas. Vet Comp Oncol. 2004:2(1):24-35.
38. Menéndez JA, Mehmi I, Griggs DW and Lupu R. The angiogenic factor CYR61 in breast cancer: Molecular pathology and therapeutic perspectives. Endocr Relat Cancer. 2003:10(2): 141-52.
39. Liu WS, FU XG, Li RM. CNN1 regulates the DKK1/Wnt/β-catenin/c-myc signaling pathway by activating TIMP2 to inhibit the invasion, migration and EMT of lung squamous cell carcinoma cells.Exp Ther Med. 2021;22(2):855.
40. Cheriyamundath SC, Basu S, Haase G, Doernberg H, Gavert N, Brabletz T, et al. ISG15 induction is required during L1-mediated colon cancer progression and metastasis. Oncotarget. 2019 ; 10(67):7122-31.
41. Desai SD. ISG15: A double edged sword in cancer. OncoImmunology. 2015; 4(12):e1052935.
42. Villarroya-Beltri C, Guerra S, Sanchez-Madrid F. ISGylation - a key to lock the cell gates for preventing the spread of threats. J Cell Sci. 2017; 130(18):2961-9.

Figures

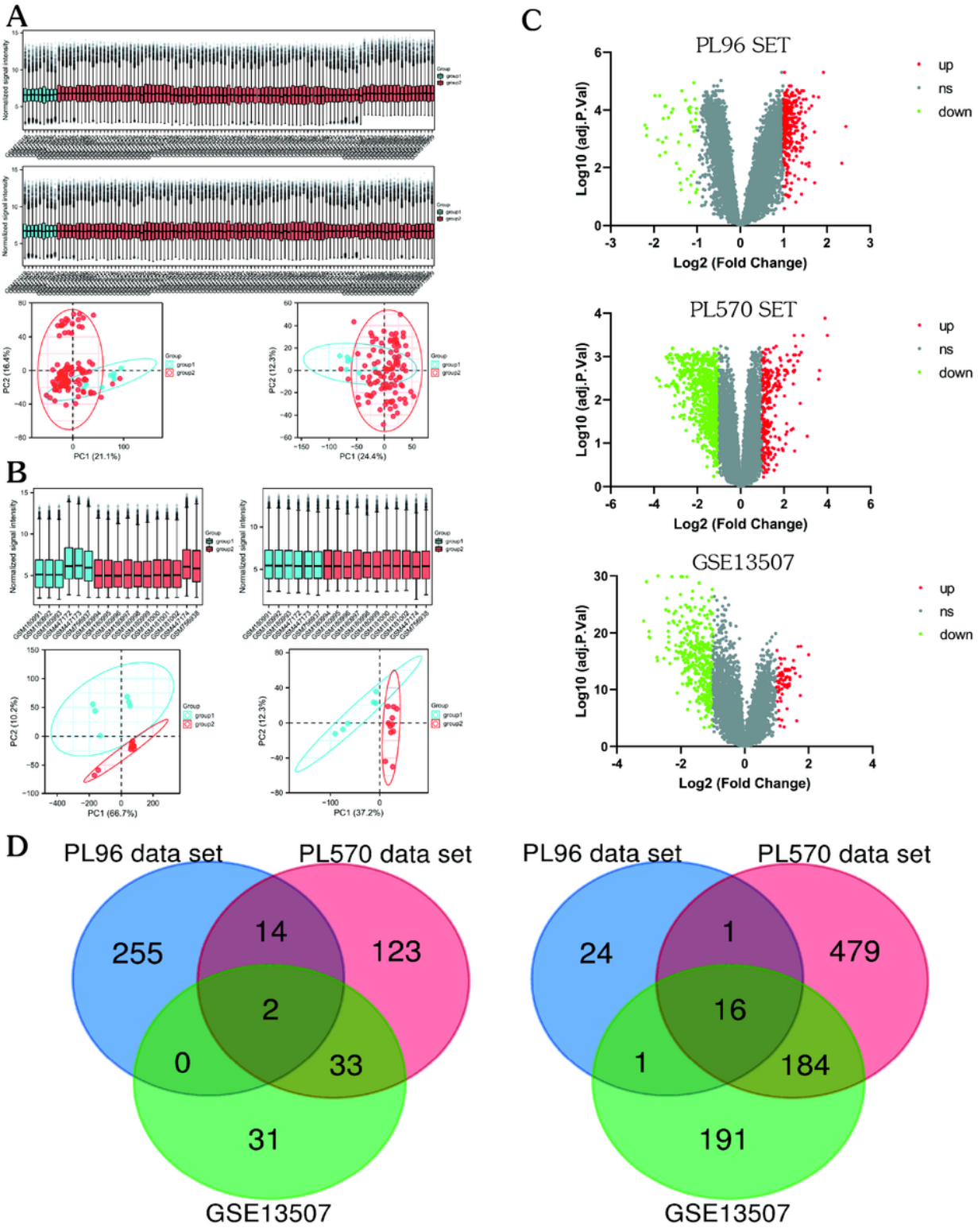

Figure 1

The common DEGs was screened by PL96 set, PL570 set and GSE13507 data. Box and PCA correction diagram before and after gene expression standardization of PL96 (A) and PL570 (B) BC data set. (C) Volcanic map of gene expression in three data sets. Red and green spots represent DEGs, red represents up-regulated genes, and green represents down-regulated genes. (D) Venn plots of up-regulated (left) and down-regulated (right) DEGs in three data sets.

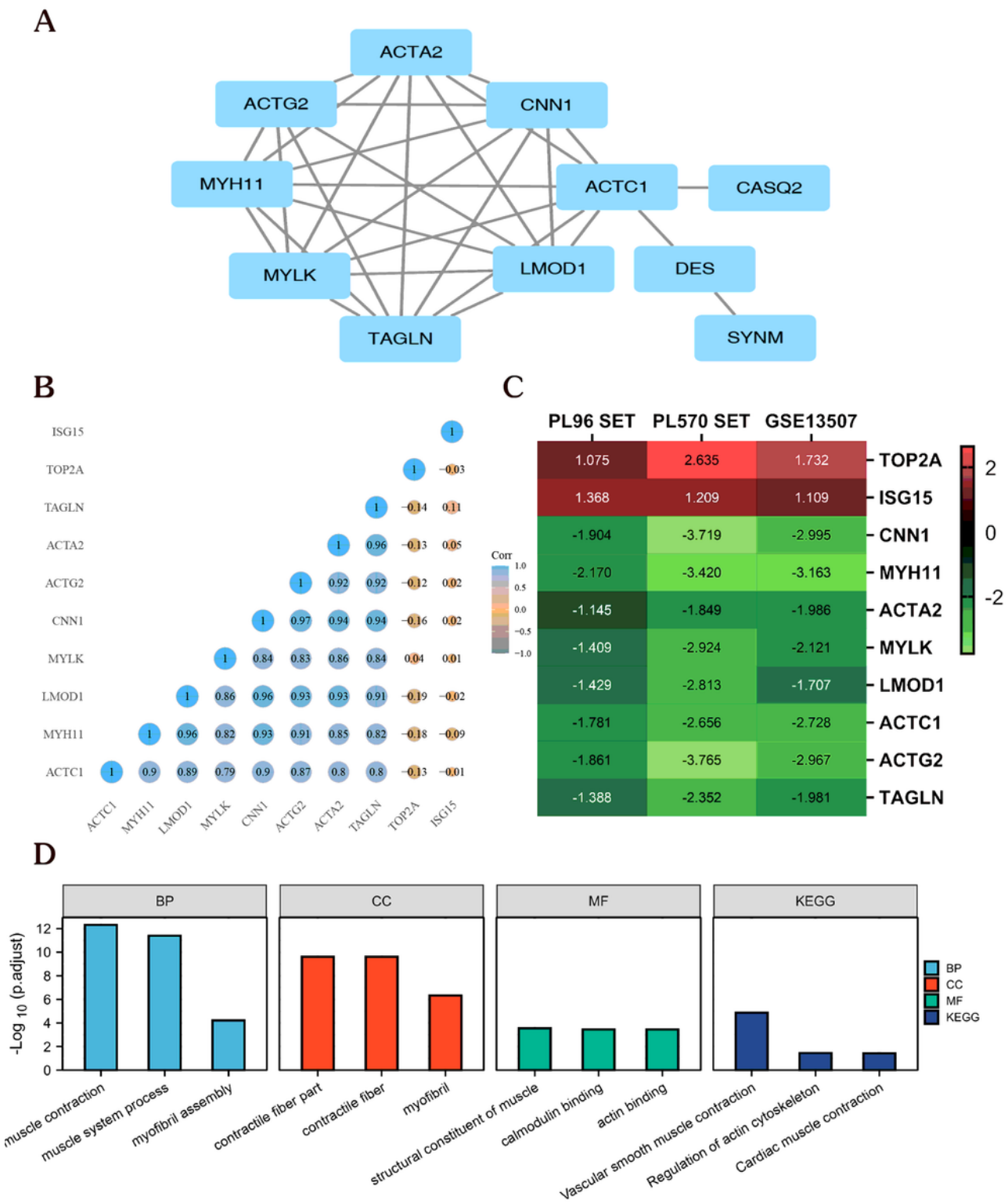

Figure 2

Correlation analysis of selected genes and Gene ontology and KEGG pathway analysis of screening down-regulated genes . (A) Interactive network diagram of down-regulated gene set derived from Cytoscape software. (B) Correlation analysis of 8 down-regulated genes and 2 up-regulated genes screened in TCGA database. (C) Differential screening gene expression heat map, in which different

colors represent the expression trend in three data sets. (D) BP, CC, MF enrichment and KEGG pathway analysis was carried out on the down-regulated DEGs.

Figure 3

Expression analysis of candidate genes in BC. (A) The expression level of selected genes in various types of tumor tissues and normal tissues in the Oncomine database. (P value is 0.001, fold change is 1.5, and gene ranking of all. The representative diagram of the expression difference about TOP2A (B) and CNN1 (C) in the Oncomine database. (D) Screened gene expression in BC and normal tissues in TCGA and GTEx cancer data displayed by GEPIA online database. In the TCGA database, the expression divergence of selected genes in different groups: (E) the expression divergence between BC and paired paracancerous tissues. (F) the expression discrepancy in paracancerous tissues, poorly differentiated and well differentiated cancers; (G) the expression difference between adjacent normal tissues and different pathological stages. (Significance identification: NS, $P \geq 0.05$; *, $p < 0.05$; **, $p<0.01$; ***, $p<0.001$.)

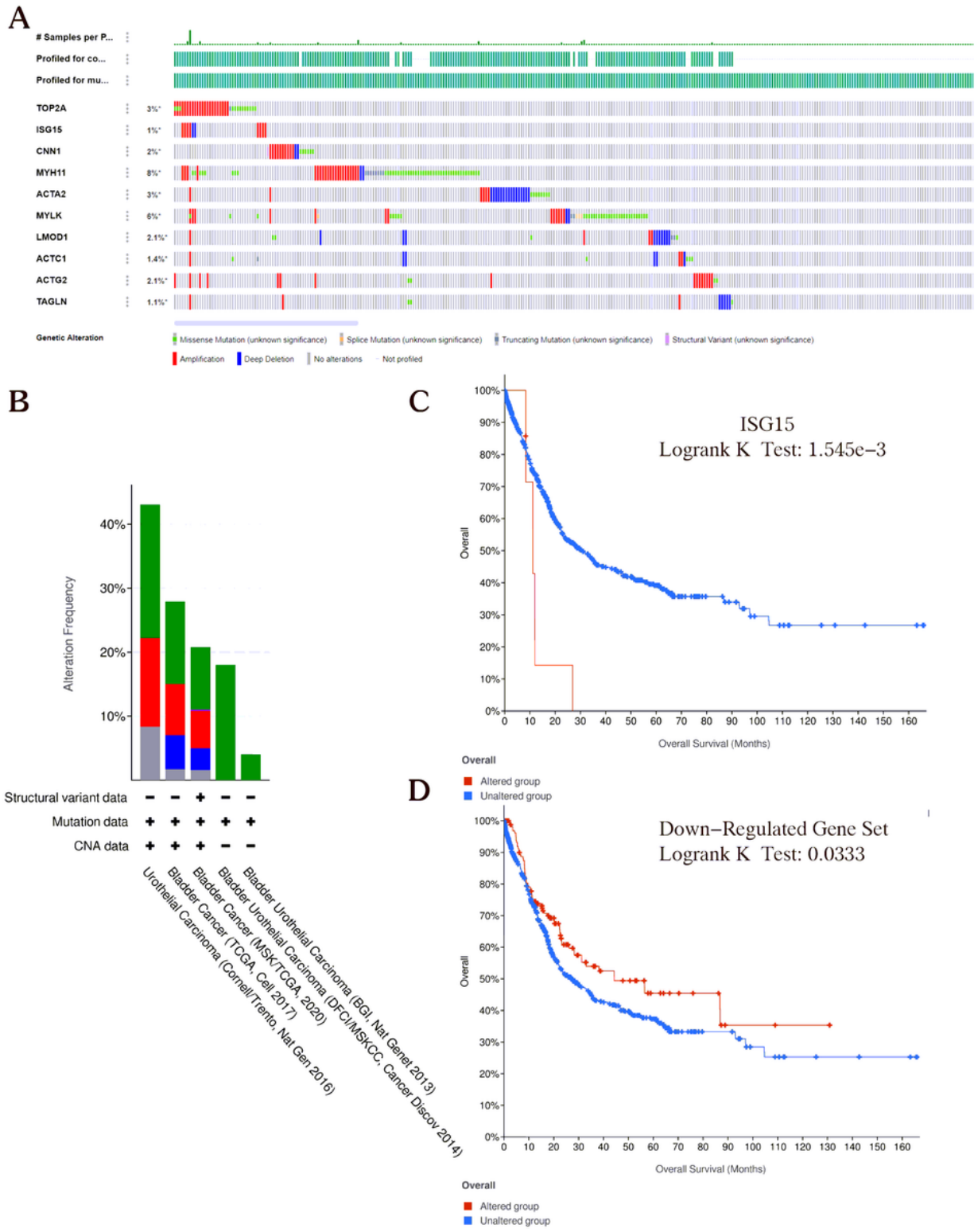

Figure 4

Mutation analysis of candidate genes in cBioPortal database. Overview map of mutations in each gene (A) and the total gene (B). Different colors represent distinct mutation types. The influence of gene mutation: the down-regulated gene set (C) and ISG15 (D) mutation altered the overall survival rate of BC patients.

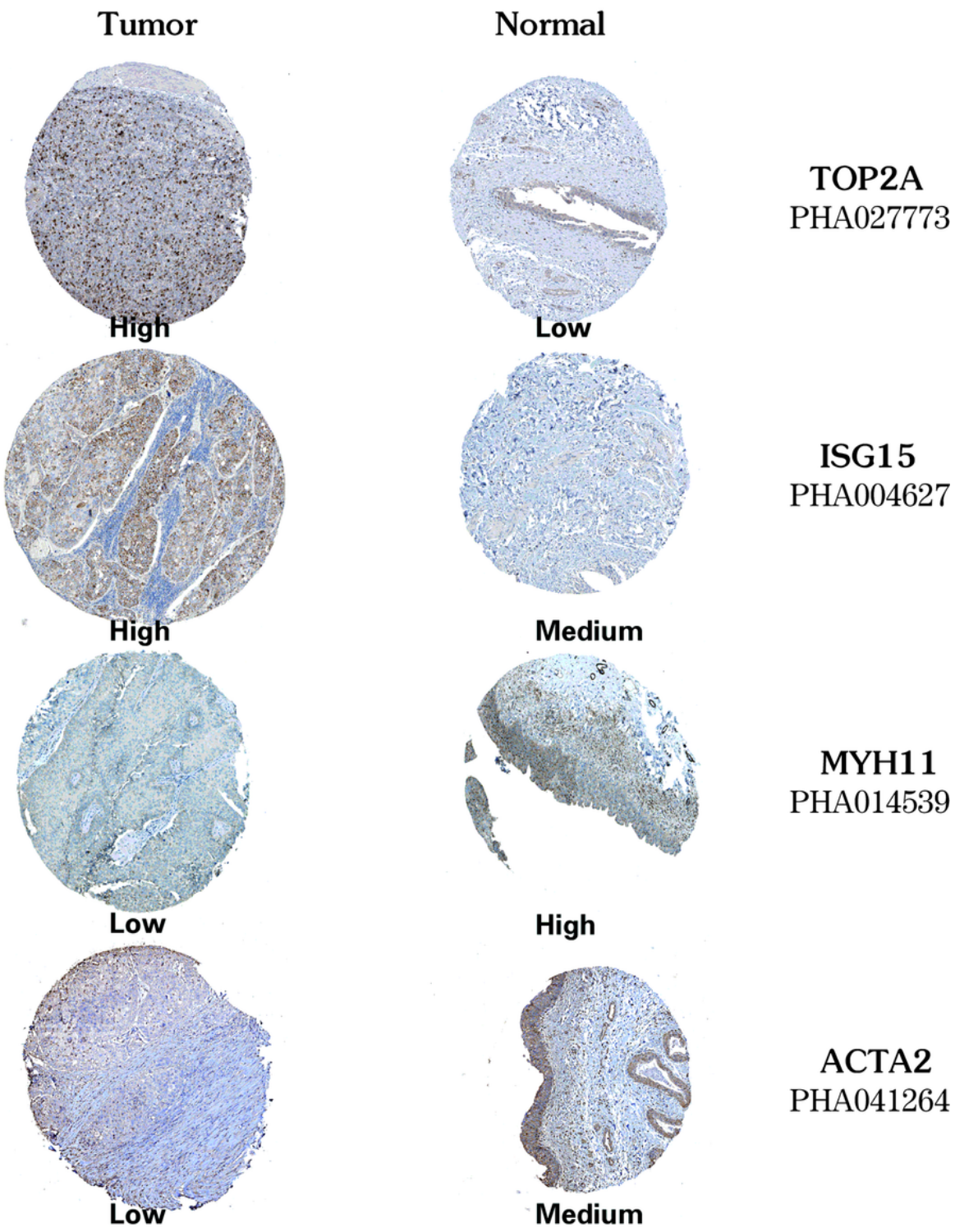

Figure 5

In the HPA database, the immunohistochemical staining of each gene in normal and cancerous tissues. The left staining picture showed the cancerous tissue and the right demonstrated the normal one. Each line corresponded to a common gene name and antibody number.

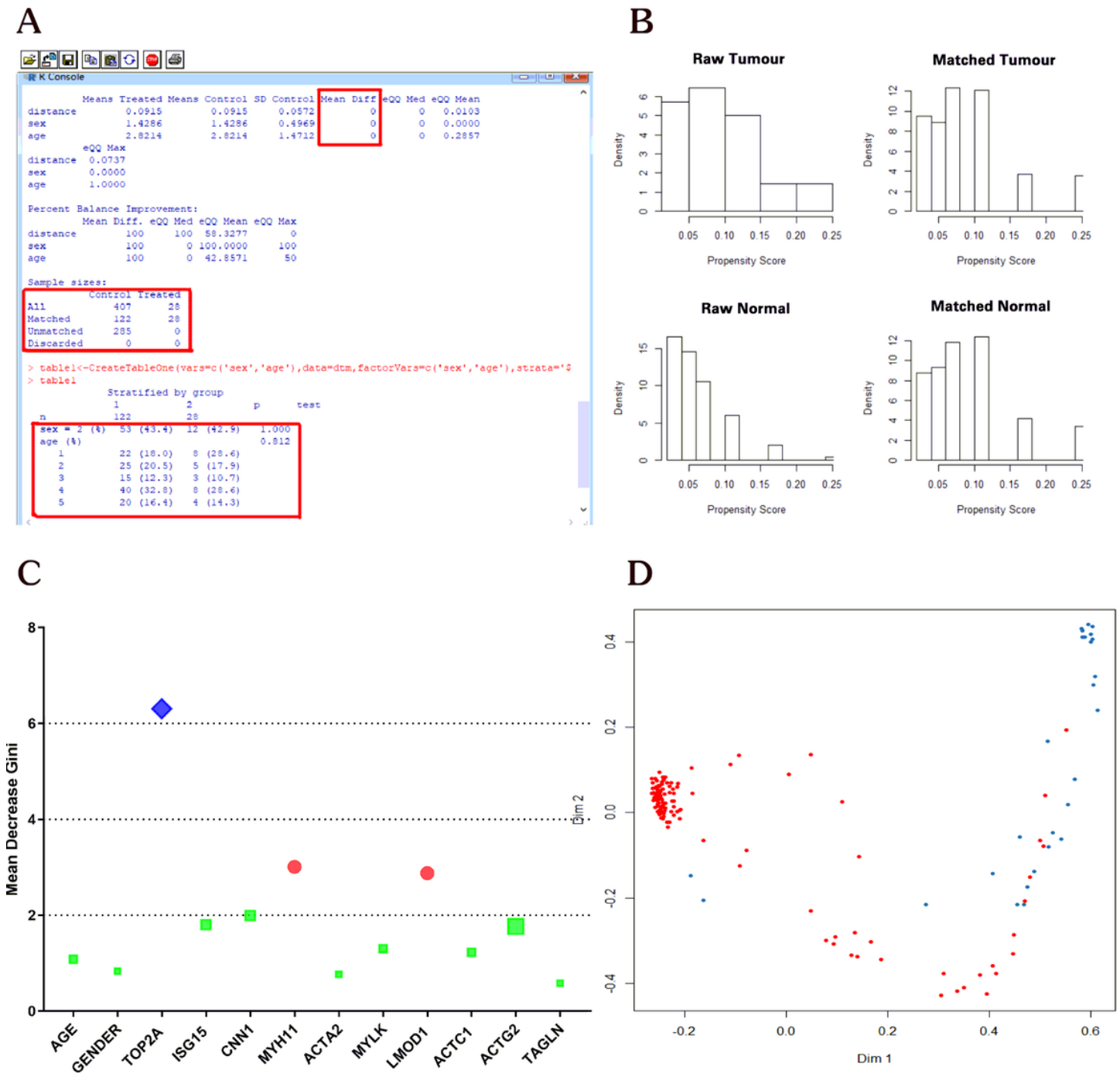

Figure 6

Random forest analysis after PSM. The results of PSM (A) for the cases collected by TCGA and GTX and the histogram before and after matching. The histogram (B) indicated that the two groups had good similarity after matching. The subsequent random forest analysis showed the MDG score of each gene (C) and the multidimensional scale map (D) of each case after scoring. The different colored dots in the figure represented the samples of the tumor group or the control group. The effectiveness of random forest analysis was evaluated by the trend of concentration within groups and dispersion between groups.

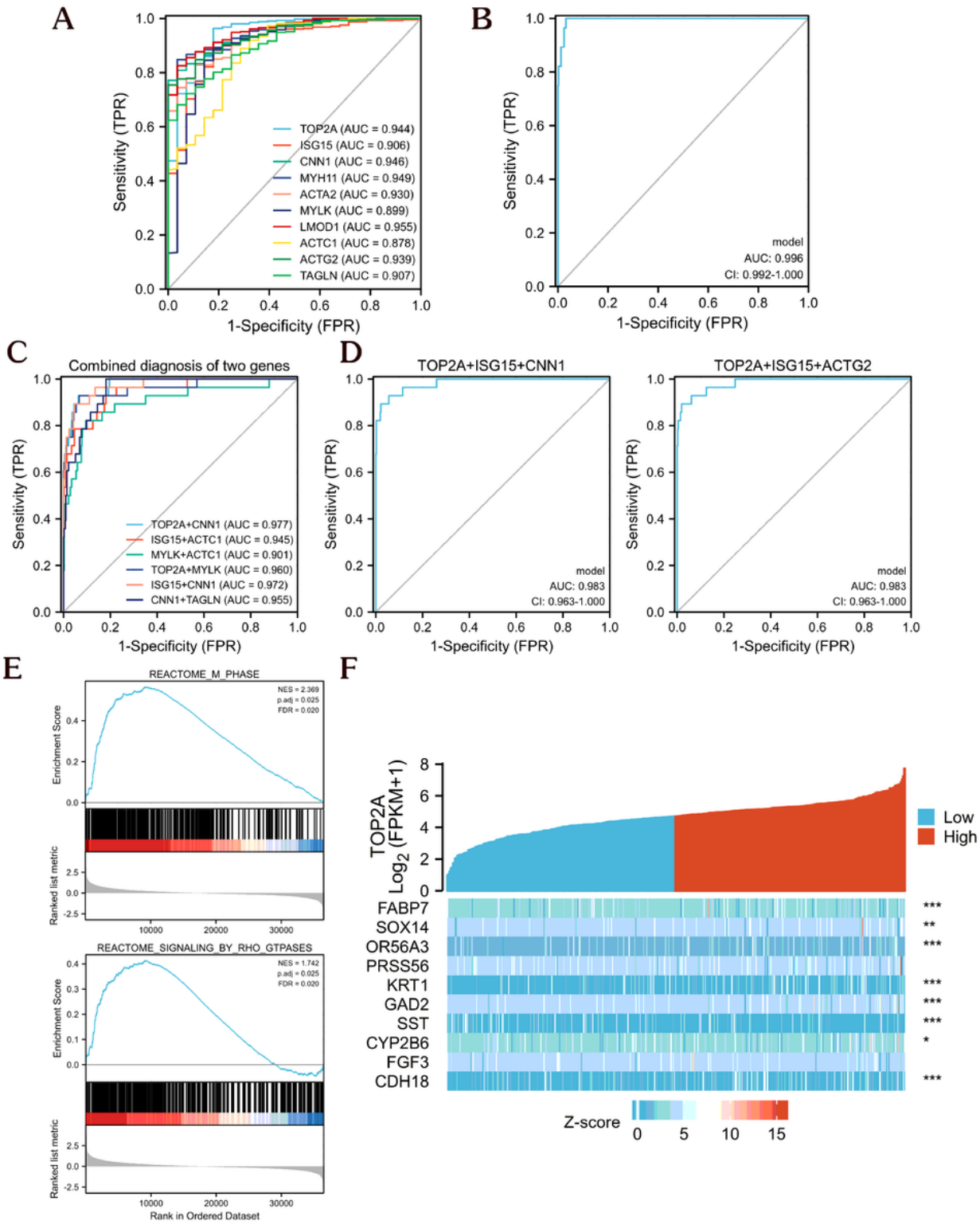

Figure 7

ROC diagnostic analysis, TOP2A GSEA and gene correlation analysis. Diagnostic value of individual (A) and combined (B) genes. Figure 7C showed the joint diagnostic efficacy of any two genes of 10 candidate genes, and the combination mode of obtaining the maximum AUC value was TOP2A add CNN1; Likewise, the maximum efficiency modes of joint diagnosis of three random genes were shown in Figure 7D, the combination were TOP2A + ISG15 + cnn1 and TOP2A + ISG15 + ACTG2 respectively. (E)

GSEA results showed by_rho_gtpases and M_phase differentially enriched in TOP2A-related BC. (F) The heat map of correlation analysis revealed 10 genes most closely related to TOP2A gene.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplymentadditionaltable.docx