Title: Identification of Cell Division Cycle 20 as a Candidate Biomarker and Potential Therapeutic Target in Bladder Cancer Using Bioinformatics Analysis

Running title: Identification of CDC20 for bladder cancer

Authors: Peilin Shen¹, Xuejun He¹, Lin Lan², Yingkai Hong¹*, Mingen Lin¹*

¹ Department of Urology, The First Affiliated Hospital of Shantou University Medical College, Shantou, China
² Shantou University Mental Health Center, Shantou, China

* Mingen Lin and Yingkai Hong should be considered joint correspondence authors

Mingen Lin
E-mail: sdfyurology@163.com
Postal address: Department of Urology, The First Affiliated Hospital of Shantou University Medical College, No. 57, Changping Road, Jinping District, Shantou, Guangdong, China

Yingkai Hong
E-mail: hyklwj@163.com
Postal address: Department of Urology, The First Affiliated Hospital of Shantou University Medical College, No. 57, Changping Road, Jinping District, Shantou, Guangdong, China
Abstract

Purpose: As Bladder Cancer (BC) is very heterogeneous and complicated in the genetic level, exploring genes to serve as biomarkers and therapeutic targets is practical.

Materials and Methods: We searched Gene Expression Omnibus (GEO) and downloaded the eligible microarray datasets. After intersection analysis for identified differentially expressed genes (DEGs) of included datasets, overlapped DEGs were identified and subsequently analyzed with Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), Protein–Protein Interaction (PPI) and hub genes identification. Hub genes were further analyzed with mRNA expression comparison in Oncomine and Gene Expression Profiling Interactive Analysis database, proteomics-based validation in The Human Protein Atlas and survival analysis in GEO and Oncolnc database.

Results: We analyzed 5 eligible GEO datasets and identified 76 overlapped DEGs mapped into PPI network with 459 edges which were mainly enriched in cell cycle pathway and related terms in GO and KEGG analysis. Among 5 identified hub genes, which are Cyclin-Dependent Kinase 1 (CDK1), Ubiquitin-Conjugating Enzyme E2 C (UBE2C), Cell Division Cycle 20 (CDC20), Microtubule Nucleation Factor (TPX2) and Cell Division Cycle Associated 8 (CDCA8); CDC20 and CDCA8 were confirmed as significant in mRNA expression comparison and proteomics-based validation. However, only CDC20 was considered prognostically significant in both GEO and Oncolnc database.

Conclusions: CDC20 and CDCA8 were identified as candidate diagnostic biomarkers for BC in the presented study; however, only CDC20 was validated as prognostically valuable and may possibly serve as a candidate prognostic biomarker and potential therapeutic target. Still, further validation studies are essential and indispensable.

1 Introduction

Bladder cancer (BC) is the 9th most common diagnostic cancer and the second most common diagnostic uro-oncological disease worldwide [1, 2]. In 2018, 81,190 people in the USA were diagnosed with BC, among which approximately 75% were non-muscle-invasive bladder cancer (NMIBC), and 25% were muscle-invasive bladder cancer (MIBC) [3, 4]. Despite there are various treatment modalities, the prognosis of BC is still far from satisfactory even after appropriate therapy. In NMIBC, the recurrence rate is about 50~70%, and the progression rate is 1~2% in low-grade tumor and nearly 45% in high-grade tumor [5-7]. After progressing to MIBC, the 5-year survival rate is less than 50% [8, 9]. As a tumor with lower grade and earlier stage usually has a better prognosis after treatment, early detection of BC is essential for the improvement of survival rate.

Even to this day, cystoscopy and urinary cytology are still the standard diagnostic and follow-up method for BC, however, both with some shortcomings. Cystoscopy is invasive and with the risk of a series of complications despite having high sensitivity and specificity. As an alternative to cystoscopy, although urinary cytology is noninvasive and highly specific, the overall sensitivity rate is only 33~48% [10, 11]. Nowadays, many novel biomarkers for BC have
emerged, but none of them has been confirmed as having reasonable sensitivity and specificity to be applied widely. Major guidelines of bladder cancer such as NCCN guidelines [12], EUA guidelines [13] and AUA guidelines [14] only recommend current biomarkers with low evidence strength. What’s more, protein markers such as bladder tumor antigen (BTA) and nuclear matrix protein 22 (NMP22) reflect mainly an infection or inflammation rather than the oncologic characteristics of the tumor which may increase the misdiagnosis rate.

BC is a very heterogeneous and complicated disease in the genetic level. As it has been reported that BC is highly associated with multiple mRNA, lncRNA and miRNA, new genetic biomarkers may provide more important information than protein markers and serve as diagnostic and prognostic indicators [15]. Furthermore, genes with prognostic value may be involved with the mechanism of tumor genesis and development which can be explored as therapeutic targets as well.

To help identify sufficient biomarker with diagnostic and prognostic value and potential therapeutic target for BC, we performed this integrated bioinformatics analysis using the data from Gene Expression Omnibus (GEO), Oncomine, Genotype-Tissue Expression (GTEx) project, The Human Protein Atlas (THPA) and The Cancer Genome Atlas (TCGA).

2 Materials & Methods

The flow chart of integrated bioinformatics analysis in the presented study is showed in Figure 1. The full R code and generated expression matrixes of every GEO dataset we used are provided as supplementary materials in Appendix 1 and 2.

Figure 1. Flow chart of bioinformatics analysis.

Legends

GEO, Gene Expression Omnibus; DEGs, differentially expressed genes; GO, Gene Ontology; PPI, Protein-Protein Interaction; KEGG, Kyoto Encyclopedia of Genes and Genomes; THPA, The Human Protein Atlas; GEPIA, Gene Expression Profiling Interactive Analysis; TCGA, The Cancer Genome Atlas.

2.1 Gene Expression Omnibus (GEO) microarray data pre-processing

We performed a search in the GEO database (www.ncbi.nlm.nih.gov/geo/) for candidate microarray datasets related to BC through September 14, 2019. The term “bladder cancer” was used as a keyword for searching, and the organism was restricted in “Homo sapiens”. After carefully screening the retrieving results including various types of samples and expression data according to the inclusion and exclusion criteria, the eligible datasets were taken to the next step of data pre-processing. The inclusion criteria are as follows: (1) human bladder tumor samples versus normal bladder tissue samples; (2) mRNA expression profile; (3) available raw data. The exclusion criteria are as follows: (1) bladder tumor samples were restricted in a single pathological type; (2) raw data were unable to be processed due to technical issue; (3) sample size was less than 10.
All eligible datasets were pre-processed individually using R language software. Robust multi-array average (RMA) method [16] was used for background correcting, normalization and summarization. We considered datasets with high degree of inconsistency in the median, the upper and lower quartiles as with poor quality which would be ruled out. Annotation was performed using different annotation packages according to different platforms of the datasets.

2.2 Integrated analysis of gene expression datasets
Eligible microarray datasets were further analyzed in R language. By using the t-test method in the Linear Models for Microarray (LIMMA) package [17], we identified differentially expressed genes (DEGs) between tumor samples and normal bladder tissue samples with the cut-off criteria of fold change≥2.0 and P<0.01. Subsequently, volcano plots were generated to demonstrate the results of DEGs selection using the ggplot2 package [18]. Finally, the intersection analysis was performed for the DEGs of all datasets using Venn diagram webtool (bioinformatics.psb.ugent.be/webtools/Venn/) to picked out the overlapped DEGs.

2.3 Gene Ontology (GO) annotation analysis and pathway enrichment analysis
We used clusterProfiler package [19] in R language to conduct Gene Ontology (GO) [20] annotation analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) [21] pathway enrichment analysis for the overlapped DEGs. Adjusted P value<0.05 and gene counts>10 were considered as statistically significant in the 3 components of GO analysis, including biological process (BP), cellular component (CC) and molecular function (MF). As for KEGG analysis for biochemistry pathways, we set the cut-off criteria as adjusted P value<0.05 to indicate a statistical significance.

2.4 Protein–Protein Interaction (PPI) network construction and hub genes identification
We established a protein–protein interaction (PPI) network by mapping the overlapped DEGs to the Search Tool for the Retrieval of Interacting Genes (STRING) database (version 10.5) (string-db.org) [22]. Interactive relationships among the overlapped DEGs were considered statistically significant with a combined score>0.4. The result generated by the STRING database was input into Cytoscape software (version 3.6.1) [23] for visualization demonstration. CytoHubba (version 0.1) [24], a plugin in Cytoscape was used to identify the hub genes which defined as the top 5 ranked genes according to the connectivity degree levels of each protein node.

2.5 Assessment of hub genes in Oncomine and Gene Expression Profiling Interactive Analysis (GEPIA) database
The identified hub genes were assessed through comparative analysis of mRNA in Oncomine (version 4.5) (www.oncomine.org) [25] and GEPIA (gepia.cancer-pku.cn/index.html) [26]. Oncomine database is a publicly accessible online data-mining platform containing various cancer microarray data and provides multiple kind of integrated analysis of gene expression. GEPIA is a web tool for RNA sequencing expression data analysis of 9,736 tumors and 8,587...
normal samples from the TCGA database and the GTEx projects. Comparative analysis of each hub gene was performed individually in 2 databases. In Oncomine, the analysis type was restricted in “Bladder Cancer vs. Normal Analysis” and the data type was restricted in “mRNA”. After the comparison of hub gene expression between BC samples and normal bladder tissue samples across included analyses, a comparison figure would be generated by the database automatically demonstrating the median rank and combined P value of each hub gene. As for GEPIA, we used box-plots to compare the mRNA expression of hub genes between TCGA bladder tumors vs TCGA normal bladder tissues + GTEx normal bladder tissues with the cut-off criteria of $|\text{Log2FC}|>1$ and $P$ value $<0.01$. Those with significant overexpression in BC in both databases were taken to further proteomics-based validation.

### 2.6 Proteomics-based validation of hub genes in The Human Protein Atlas (THPA)

THPA [27] contains immunohistochemistry images showing antibody staining in samples from 144 individuals corresponding to 44 different normal tissue types, and samples from 216 cancer patients corresponding to 20 different types of cancer by using tissue microarrays technology. We downloaded the histological section images and corresponding information of significantly overexpressed hub genes from normal urinary bladder tissues and urothelial cancer tissues of bladder obtained by immunohistochemistry in THPA. Because the antibody staining is already reported as not detected, low, medium, or high based on the staining intensity and fraction of stained cells, we conducted the Mann-Whitney Test in SPSS 19.0 to compare the antibody staining level of hub genes between normal urinary bladder tissues and urothelial cancer tissues of bladder. The cutoff P value was set as 0.05 and the hub genes with significant antibody staining in urothelial cancer cells were considered as significant hub genes which may be useful for BC diagnosis.

### 2.7 Survival analysis of hub genes using Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) data

In order to determine the association between hub genes and clinical outcomes, significant hub genes validated by mRNA comparative analysis and proteomics-based validation were taken to further survival analysis. Expression and survival data from GEO were analyzed in R language software with the survival package [28], while survival plots were automatically generated in the online database Oncolnc (www.oncolnc.org) which is based on TCGA data and contains both RNA expression data and survival data of up to 21 different kinds of cancers [29]. A hub gene with $P<0.05$ was considered as a prognostically valuable gene, the high expression of which was significantly related to poor prognosis; therefore, it can be regarded as a candidate prognostic biomarker or potential therapeutic target for BC.

### 3 Results

#### 3.1 Identification of differentially expressed genes (DEGs) in BC
With the search of GEO database, 4795 results were found for screening, in which 7 microarray datasets including GSE7476, GSE13507, GSE31189, GSE40355, GSE52519, GSE65635 and GSE121711 appeared to be eligible according to inclusion and excluding criteria. After preprocessing, the results of all datasets showed superior quality except for the result of GSE31189 and GSE121711, and therefore excluded for further analysis (Appendix 3). The 5 remaining datasets are all with a sample size >10 and the platforms utilized by different datasets were various (Table 1). By screening the expressed genes with the cut-off criteria, the DEGs were identified for each dataset. The up-regulated DEGs were 425 in GSE7476, 182 in GSE13507, 2008 in GSE40355, 211 in GSE52519 and 1470 in GSE65635. The down-regulated DEGs were 1233 in GSE7476, 634 in GSE13507, 2510 in GSE40355, 315 in GSE52519 and 1761 in GSE65635 (Fig.2). Subsequently, 29 up-regulated and 47 down-regulated overlapped DEGs were calculated using intersection analysis (Fig.3).

Table 1. Characteristics of included datasets

| Dataset   | Platform | Sample Size |
|-----------|----------|-------------|
| GSE7476   |          | >10         |
| GSE13507  |          | >10         |
| GSE31189  |          | >10         |
| GSE40355  |          | >10         |
| GSE52519  |          | >10         |
| GSE65635  |          | >10         |

Figure 2. Identified DEGs of included datasets.

Legends

DEGs between BC samples and normal bladder tissue samples were identified with the cut-off criteria of fold change ≥2.0 and P<0.01. The up-regulated DEGs were 425 in GSE7476, 182 in GSE13507, 2008 in GSE40355, 211 in GSE52519 and 1470 in GSE65635. The down-regulated DEGs were 1233 in GSE7476, 634 in GSE13507, 2510 in GSE40355, 315 in GSE52519 and 1761 in GSE65635.

Figure 3. Overlapped DEGs from the intersection analysis of included datasets.

Legends

DEGs of GSE7476, GSE13507, GSE40355, GSE52519 and GSE65635 were input into Venn diagram webtool for intersection analysis and 76 overlapped DEGs with 29 up-regulated genes and 47 down-regulated genes were identified.

3.2 Functional enrichment analysis

All 76 overlapped DEGs were processed with GO and KEGG analysis. For GO analysis, BP was the most favorable enrichment component, in which DEGs were significantly enriched in mitotic nuclear division, nuclear division, organelle fission and so on (Fig.4a). CC analysis was enriched in contractile fiber part, contractile fiber, myofibril, et al (Fig.4b). For MF analysis, although there were significant enriched terms, none of them contains more than 10 DEGs. In addition, the results of KEGG pathway enrichment analysis indicated that DEGs were most enriched in cell cycle pathway (Fig.4c).

Figure 4. GO annotation analysis and KEGG Pathway enrichment analysis of overlapped DEGs.

Legends
Terms enriched in biological processes of GO enrichment analysis were as follow: mitotic nuclear division, nuclear division, organelle fission, muscle system process, muscle contraction, regulation of mitotic nuclear division, sister chromatid segregation, regulation of nuclear division, regulation of mitotic cell cycle phase transition, chromosome segregation, nuclear chromosome segregation, regulation of cell cycle phase transition, muscle cell differentiation. (b) Terms enriched in cellular components of GO enrichment analysis were as follow: contractile fiber part, contractile fiber, myofibril, midbody, sarcomere, spindle, actin cytoskeleton. (c) Pathways enriched in KEGG analysis were Cell cycle, Hypertrophic cardiomyopathy (HCM), Dilated cardiomyopathy (DCM), Progesterone-mediated oocyte maturation, Focal adhesion and Oocyte meiosis.

3.3 PPI network and hub genes

PPI network was constructed with STRING tool and visualized with Cytoscape software. There were 76 nodes and 459 edges in the PPI network generated by STRING tool originally. After removing 18 unconnected nodes, the network with 58 nodes and 459 edges was visualized in Cytoscape, which roughly clustered into two sets depending on the types of genes were up-regulated or down-regulated (Fig.5). In addition, the PPI enrichment P value is less than 1.0E-16 which reflected significant interactions among the overlapped DEGs. The calculated top 5 hub genes according to connectivity degree were all up-regulated genes and ranked in the sequence of Cyclin-Dependent Kinase 1 (CDK1), Ubiquitin-Conjugating Enzyme E2 C (UBE2C), Cell Division Cycle 20 (CDC20), Microtubule Nucleation Factor (TPX2) and Cell Division Cycle Associated 8 (CDCA8) (Table 2).

Figure 5. PPI network of the overlapped DEGs.

Legends

The PPI network contains 58 nodes and 459 edges which roughly clustered into two sets depending on the types of genes were up-regulated or down-regulated.

Table 2. Top five hub genes with the highest degree of connectivity

3.4 Comparative mRNA expression analysis of hub genes

Identified hub genes were assessed with comparative pooled analysis using Oncomine and GEPIA database. All 5 hub genes were significantly overexpressed in BC samples in GEPIA database (Fig.6). In Oncomine database, only TPX2 (P=0.171) showed a statistically insignificant difference when comparing the gene expression between BC sample and normal bladder tissue samples while CDK1 (P=0.015), UBE2C (P=9.29E-7), CDC20 (P=2.07E-8) and CDCA8 (P=1.22E-13) were validated as significant (Fig.7). CDK1, UBE2C, CDC20 and CDCA8 showing significant higher expression in BC samples in both databases were taken to further proteomics-based validation.
Figure 6. Box-plots of hub genes mRNA expression between bladder cancer and normal bladder in GEPIA database.

**Legends**

We compared the mRNA expression of CDK1, UBE2C, CDC20, TPX2 and CDCA8 between bladder cancer and normal bladder in GEPIA database using data from TCGA and GTEx. We considered |Log2FC|>1 and P value<0.01 as significant and all 5 genes met the statistical standard.

Figure 7. Pooled comparative analyses on the mRNA expression of hub genes in Oncomine database.

**Legends**

* The rank for a gene is the median rank for that gene across each of the analyses.
† The p-Value for a gene is its p-Value for the median-ranked analysis.

1 Infiltrating Bladder Urothelial Carcinoma vs. Normal (Blaveri Bladder 2, Clin Cancer Res, 2005)
2 Superficial Bladder Cancer vs. Normal (Blaveri Bladder 2, Clin Cancer Res, 2005)
3 Infiltrating Bladder Urothelial Carcinoma vs. Normal (Dyrskjot Bladder 3, Cancer Res, 2004)
4 Stage 0is Bladder Urothelial Carcinoma vs. Normal (Dyrskjot Bladder 3, Cancer Res, 2004)
5 Superficial Bladder Cancer vs. Normal (Dyrskjot Bladder 3, Cancer Res, 2004)
6 Infiltrating Bladder Urothelial Carcinoma vs. Normal (Lee Bladder, J Clin Oncol, 2010)
7 Superficial Bladder Cancer vs. Normal (Lee Bladder, J Clin Oncol, 2010)
8 Infiltrating Bladder Urothelial Carcinoma vs. Normal (Modlich Bladder, Clin Cancer Res, 2004)
9 Superficial Bladder Cancer vs. Normal (Modlich Bladder, Clin Cancer Res, 2004)
10 Infiltrating Bladder Urothelial Carcinoma vs. Normal (Sanchez-Carbaya Bladder 2, J Clin Oncol, 2006)
11 Superficial Bladder Cancer vs. Normal (Sanchez-Carbaya Bladder 2, J Clin Oncol, 2006)

3.5 Proteomics-based validation of hub genes

We used THPA database to further verify the significance of CDK1, UBE2C, CDC20 and CDCA8 in protein level with immunohistochemistry images. High antibody staining images of all 4 genes can be found on THPA (Fig.8a) and high positive staining rate was detected in urothelial cancer tissues of bladder for CDK1 (44/46), UBE2C (57/57), CDC20 (43/47), and CDCA8 (57/61). However, only CDC20 (P=0.006) and CDCA8 (P=0.007) showed statistically significant results by comparing normal urinary bladder with urothelial cancer tissues of bladder using Mann-Whitney Test, and can be considered as significant hub genes which may be useful for BC diagnosis. (Fig.8b)

Figure 8. Antibody staining of hub genes in immunohistochemistry images between normal urinary bladder and urothelial cancer tissues of bladder.

**Legends**
(a) High antibody staining immunohistochemistry images of 4 hub genes are available at https://www.proteinatlas.org/ENSG00000170312-CDK1/pathology/urothelial+cancer#img (CDK1), https://www.proteinatlas.org/ENSG00000175063-UBE2C/pathology/urothelial+cancer#img (UBE2C), https://www.proteinatlas.org/ENSG00000117399-CDC20/pathology/urothelial+cancer#img (CDC20) and https://www.proteinatlas.org/ENSG00000134690-CDCA8/pathology/urothelial+cancer#img (CDCA8) in The Human Protein Atlas, respectively.

(b) 2 antibodies were used for CDK1 (HPA003387 and CAB003799) and CDC20 (CAB004525 and HPA055288) and 3 for UBE2C (CAB011464, CAB035990 and HPA054975) and CDCA8 (HPA028120, HPA028258 and HPA028783). High positive staining rate was detected in urothelial cancer tissues of bladder for CDK1 (44/46), UBE2C (57/57), CDC20 (43/47), and CDCA8 (57/61), however, only CDC20 (P=0.006) and CDCA8 (P=0.007) showed significant results by comparing normal urinary bladder with urothelial cancer tissues of bladder using Mann-Whitney Test. The cutoff P value was set as 0.05.

3.6 Survival analysis of significant hub genes

The survival analysis was conducted for 2 significant hub genes. According to the survival plots in both 2 databases, high expression of CDC20 level was significantly associated with poor overall survival (P=0.001 in GEO; P=0.0361 in Oncolnc), which suggested that CDC20 may contribute to the mortality of BC patients. On the other hand, CDCA8 only showed a statistical significance in GEO (P<0.001 in GEO; P=0.277 in Oncolnc) which indicated that it is less possible to be further explored as a candidate prognostic biomarker or potential therapeutic target for BC (Fig.9).

Figure 9. Survival analysis of significant hub genes.

Legends

We conducted survival analysis using GEO data with R language software and TCGA data in Oncolnc database to demonstrate the association between the expression of significant hub genes and overall survival time. CDC20 showed a statistical significance in both database (P=0.001 in GEO; P=0.0361 in Oncolnc), while CDCA8 only showed a statistical significance in GEO (P<0.001 in GEO; P=0.277 in Oncolnc).

4 Discussion

As one of the most common diagnostic urologic tumors, it has been reported that BC is with great heterogeneity and complexity at the molecular level. Therefore, through exploring deep into the genetic level with integrated bioinformatics analysis, we tried to discover critical genes which can act as novel biomarkers with diagnostic and prognostic value or potential therapeutic targets for BC.

We searched the GEO database and selected 5 datasets for inclusion. By using R language software and Venn diagram webtool, 76 DEGs with 29 up-regulated genes and 47 down-
regulated genes were identified. After GO annotation analysis and KEGG pathway enrichment analysis, DEGs were mostly enriched in terms and pathways related to cell cycle which play an important role in cell growth, anabolism and proliferation [30]. Precise cell proliferation and DNA replication are essential for maintaining genomic stability and uncontrolled cell cycle procedure is closely associated with tumor genesis and development, including BC [31, 32].

Then, a PPI network with 76 nodes and 459 edges were constructed to better demonstrate the connection among identified DEGs and 5 hub genes were found with top connectivity degree, including CDK1, UBE2C, CDC20, TPX2 and CDCA8. The expression of each hub gene between BC and normal bladder tissues was assessed in Oncomine and GEPIA database individually. Except for TPX2, 4 other hub genes with significant overexpression in BC sample in both Oncomine and GEPIA were taken to further proteomics-based validation. By analyzing the antibody staining immunohistochemistry images of CDK1, UBE2C, CDC20 and CDCA8 in THPA using Mann-Whitney Test, we considered CDC20 and CDCA8 as significant hub genes with diagnostic value for BC. To further analyze the value of significant hub genes, survival plots were generated using GEO and Oncomine database. Unfortunately, only CDC20 showed a significant P value in both databases and was regarded as a cancer-promoting gene which can be further explored as a prognostic biomarker or therapeutic target for BC.

CDCA8, a part of the chromosomal passenger complex (CPC), plays an important role in the cell division cycle in cancer cells [33]. There have been studies argued that CDCA8 is a required initiating factor for the oncogenesis and progression of tumors [34]. However, researches about its role in BC is still limited and there is only 1 study shows that the overexpression of CDCA8 was found in BC and correlated with poor clinicopathological features of BC patients [35].

CDC20 is the only identified hub gene significantly associated with the diagnosis and prognosis of BC in the presented study. In cell cycle progression, CDC20 encodes a protein functions as an essential regulator for cell division, the most important function of which is to combine with anaphase promoting complex (APC/C), and subsequently regulate the degradation of securin. Securin destruction promotes the degradation of cohesion, the separation of sister chromatid, and subsequently the transition from G2/M phase to G1 phase. In this case, normal cell cycle can be maintained. The dysfunction of CDC20 can contribute to poor differentiation, tumor aneuploidy and poor prognosis in multiple cancer, including BC [36]. Besides, it has been revealed that its overexpression can decrease overall survival time and recurrence-free survival time in BC patients, and even lead to progression [36]. Even so, its molecular mechanism is still not very clear, and researches concerning its association with BC is very limited. Therefore, CDC20 is with great potential to be further explored as a candidate biomarker and potential cancer therapeutic target in BC [37].

As an integrated bioinformatics analysis study, several limitations were inevitable. First, the results of this study were calculated through multiple statistical steps. Therefore, by setting different cut-off criteria, applying different statistical method or using different tools, the results would be various correspondingly. Second, the inevitable heterogeneity among different datasets would impact on the reliability of the presented study. Third, no in vitro or in vivo experiment...
were designed for validation so that the conclusion is still theoretical and further experimental test confirmations are absolutely necessary.

In conclusion, 2 multiple-cancer-associated genes including CDC20 and CDCA8 were identified as candidate diagnostic biomarkers for BC by analyzing GEO, Oncomine, GTEx, THPA and TCGA data. However, only CDC20 was validated as having prognostic value and may even serve as a candidate prognostic biomarker and potential therapeutic target for BC. For CDC20 and CDCA8, either as a biomarker or a therapeutic target, further validation research is still indispensable to confirm their clinical effect in the future, and we look forward to seeing more bioinformatical and experimental studies with larger sample size and more detailed clinical information to be carried out for the remediation and extension of our study. Hopefully, by exploring deep into CDC20, it can help move the diagnosis and therapy process of BC closer to consummation.

Appendix 1. The full R code of 5 eligible GEO datasets

Legends
We used R language software for analysis. 5 eligible datasets were pre-processed individually using RMA method for background correcting, normalization and summarization. Box-plots were drawn to identify and rule out poor quality datasets. Annotation was performed using different annotation packages according to different platforms of the datasets. Then, we used the t-test method in the LIMMA package to identified DEGs between tumor samples and normal bladder tissue samples with the cut-off criteria of fold change $\geq 2.0$ and $P<0.01$. Subsequently, volcano plots were generated to demonstrate the results of DEGs selection using the ggplot2 package. Survival analysis of GSE13507 was performed using survival package.

Appendix 2. The generated expression matrixes of 5 eligible GEO datasets

Legends
Using RMA method for pre-processing, LIMMA package for differential expression analysis and different annotation packages accordingly for gene annotation, the expression matrixes of 5 GEO datasets were generated in R language software.

Appendix 3. Box-plots of 7 GEO datasets

Legends
After background correcting, normalization and summarization, the median, lower and upper quartiles of gene expression value of all samples showed high consistencies which indicated superior qualities in GSE7476, GSE13507, GSE40355, GSE52519 and GSE65635 respectively. As in GSE31189 and GSE121711, an inferior quality was indicated due to the low consistencies of the median, lower and upper quartiles.

Acknowledgements
Funding: This work was supported by the Guangdong Medical Research Foundation (A2018103 and 201711795717822), Shantou Science and Technology Project (2017026) and Natural Science Foundation of Guangdong Province (No. 2015A030310078).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Data Availability Statement: The data analyzed for this study can be found in the GEO database (www.ncbi.nlm.nih.gov/geo/), Oncomine database (www.oncomine.org), GEPIA database (gepia.cancer-pku.cn/index.html), THPA database (www.proteinatlas.org) and Oncolnc database (www.oncolnc.org). The full R code and generated expression matrixes of every GEO dataset we used are provided as supplementary materials in Appendix 1 and 2.

Author Contributions: ME and YK designed this study. XJ and L completed the search and downloaded the data. PL and XJ analyzed the data and designed the illustrations. PL wrote the manuscript under the instruction of ME and YK. ME and YK have contributed equally to this work. All authors read and approved the final manuscript.

References
1. Stewart B, Wild C: World cancer report 2014. WHO Press, Lyon.; International Agency for Research on Cancer; 2014.
2. Siegel RL, Miller KD, Jemal A: Cancer statistics, 2016. Ca A Cancer Journal for Clinicians 2015, 60(5):277-300.
3. Cancer Stat Facts: Bladder Cancer [https://seer.cancer.gov/statfacts/html/urinb.html]
4. Aldousari S, Kassouf W: Update on the management of non-muscle invasive bladder cancer. Canadian Urological Association journal = Journal de l'Association des urologues du Canada 2010, 4(1):56-64.
5. Herr HW: Tumor progression and survival of patients with high grade, noninvasive papillary (TaG3) bladder tumors: 15-year outcome. Journal of Urology 2000, 163(1):60-62.
6. Sylvestre RJ, van der Meijden AP, Oosterlinck W, Witjes JA, Bouffioux C, Denis L, Newling DW, Kuth K: Predicting recurrence and progression in individual patients with stage Ta T1 bladder cancer using EORTC risk tables: a combined analysis of 2596 patients from seven EORTC trials. European urology 2006, 49(3):466-465; discussion 475-467.
7. Soloway MS: Bladder cancer: Lack of progress in bladder cancer--what are the obstacles? Nature Reviews Urology 2013, 10(1):5-6.
8. Türkölmez K, Tokgöz H, Reşorlu B, Köse K, Bedük Y: Muscle-invasive bladder cancer: predictive factors and prognostic difference between primary and progressing tumors. Urology 2007, 70(3):477-481.
9. Yafi FA, Brimo F, Steinberg J, Aprikian AG, Tanguay S, Kassouf W: Prospective analysis of sensitivity and specificity of urinary cytology and other urinary biomarkers for bladder cancer. Urologic oncology 2015, 33(2):66.e25-66.e31.
10. Rosser CJ, Chang M, Dai Y, Ross S, Mengual L, Alcaraz A, Goodison S: Urinary protein biomarker panel for the detection of recurrent bladder cancer. Cancer Epidemiol Biomarkers Prev 2014, 23(7):1340-1345.
12. NCCN Guidelines Version 3. 2018 Bladder Cancer
[https://www.nccn.org/store/login/login.aspx?ReturnURL=https://www.nccn.org/professionals/physician_gls/pdf/bladder.pdf]

13. Babjuk M, Burger M, Compérat E, Gontero P, Mostafid H, Palou J, van Rhijn B, Rouprêt M, Shariat S, Sylvester R et al: EAU guidelines on non-muscle-invasive bladder cancer(TaT1 and CIS). 2017.

14. Chang SS, Boorjian SA, Chou R, Clark PE, Daneshmand S, Konety BR, Pruthi R, Quale DZ, Ritch CR, Seigle JD et al: Diagnosis and Treatment of Non-Muscle Invasive Bladder Cancer: AUA/SUO Guideline. The Journal of urology 2016, 196(4):1021-1029.

15. Wieczorek E, Reszka E: mRNA, microRNA and IncRNA as novel bladder tumor markers. Clinica chimica acta; international journal of clinical chemistry 2018, 477:141-153.

16. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics (Oxford, England) 2003, 4(2):249-264.

17. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research 2015, 43(7):e47.

18. Wickham H: ggplot2: Elegant Graphics for Data Analysis: Springer Publishing Company, Incorporated; 2009.

19. Yu G, Wang L, Han Y, He Q: clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. Omics-a Journal of Integrative Biology 2012, 16(5):284-287.

20. Consortium GO: The Gene Ontology (GO) project in 2006. Nucleic acids research 2006, 34(Database issue):322-326.

21. H O, S G, K S, W F, H B, M K: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic acids research 2012, 41(Database issue):D680-D686.

22. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P et al: The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic acids research 2011, 39(Database issue):D561-568.

23. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research 2003, 13(11):2498-2504.

24. Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY: cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC systems biology 2014, 8(Suppl 4):S11.

25. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia 2004, 6(1):1-6.

26. Tang Z, Li C, Kang B, Ge G, Cheng L, Zhang Z: GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic acids research 2017(W1):W1.

27. A pathology atlas of the human cancer transcriptome. Science (New York, NY) 2017, 357(6352):660-660.

28. Therneau T: A package for survival analysis. R package version 3.1-11; 2020.

29. Anaya J: OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs. PeerJ Computer Science 2016, 2:e67.

30. Maclachlan TK, Sang N, Giordano A: Cyclins, cyclin-dependent kinases and cdk inhibitors: implications in cell cycle control and cancer. Critical reviews in eukaryotic gene expression 1995, 5(2):127.
31. Hanahan D, Weinberg RA: Hallmarks of cancer: the next generation. Cell 2011, 144(5):646-674.
32. Solomon JP, Hansel DE: The Emerging Molecular Landscape of Urothelial Carcinoma. Surgical pathology clinics 2016, 9(3):391.
33. Wang Y, Zhao Z, Bao X, Fang Y, Ni P, Chen Q, Zhang W, Deng A: Borealin/Dasra B is overexpressed in colorectal cancers and contributes to proliferation of cancer cells. Medical Oncology 2014, 31(11):1-6.
34. Dai C, Miao CX, Xu XM, Liu LJ, Gu YF, Zhou D, Chen LS, Lin G, Lu GX: Transcriptional Activation of Human CDCA8 Gene Regulated by Transcription Factor NF-Y in Embryonic Stem Cells and Cancer Cells. Journal of Biological Chemistry 2015, 290(37):22423.
35. Bi Y, Chen S, Jiang J, Yao J, Wang G, Zhou Q, Li S: CDC8 expression and its clinical relevance in patients with bladder cancer. Medicine 2018, 97(34):e11899-e11899.
36. Choi JW, Kim Y, Lee JH, Kim YS: High expression of spindle assembly checkpoint proteins CDC20 and MAD2 is associated with poor prognosis in urothelial bladder cancer. Virchows Archiv 2013, 463(5):681-687.
37. Kidokoro T, Tanikawa C, Furukawa Y, Katagiri T, Nakamura Y, Matsuda K: CDC20, a potential cancer therapeutic target, is negatively regulated by p53. Oncogene 2008, 27(11):1562-1571.
5 microarray datasets selected from GEO

data pre-processing and integrated analysis

Overlapped DEGs

GO annotation analysis
PPI network construction
KEGG pathway enrichment analysis

hub genes identification

Hub genes

Proteomics-based validation in THPA

mRNA expression comparison in Oncomine and GEPIA
Survival analysis with GEO and TCGA data

Candidate biomarker and potential therapeutic target genes

Legends

GEO, Gene Expression Omnibus; DEGs, differentially expressed genes; GO, Gene Ontology; PPI, Protein–Protein Interaction; KEGG, Kyoto Encyclopedia of Genes and Genomes; THPA, The Human Protein Atlas; GEPIA, Gene Expression Profiling Interactive Analysis; TCGA, The Cancer Genome Atlas.
DEGs between BC samples and normal bladder tissue samples were identified with the cut-off criteria of fold change ≥2.0 and P<0.01. The up-regulated DEGs were 425 in GSE7476, 182 in GSE13507, 208 in GSE40355, 211 in GSE52519 and 1470 in GSE65635. The down-regulated DEGs were 1233 in GSE7476, 634 in GSE13507, 2510 in GSE40355, 315 in GSE52519 and 1761 in GSE65635.
Regulated genes and 47 down-regulated genes were identified. Venn diagram webtool for intersection analyses and 76 overlapped DEGs with 29 up-DEGs of GSE7476, GSE13507, GSE40355; GSE52519 and GSE6635 were input into Legends.
Legends

(a) Terms enriched in biological processes of GO enrichment analysis were as follow: mitotic nuclear division, nuclear division, organelle fission, muscle system process, muscle contraction, regulation of mitotic nuclear division, sister chromatid segregation, regulation of nuclear division, regulation of mitotic cell cycle phase transition, chromosome segregation, nuclear chromosome segregation, regulation of cell cycle phase transition, muscle cell differentiation. (b) Terms enriched in cellular components of GO enrichment analysis were as follow: contractile fiber part, contractile fiber, myofibril, midbody, sarcomere, spindle, actin cytoskeleton. (c) Pathways enriched in KEGG analysis were Cell cycle, Hypertrophic cardiomyopathy (HCM), Dilated cardiomyopathy (DCM), Progesterone-mediated oocyte maturation, Focal adhesion and Oocyte meiosis.
Legends

The PPI network contains 58 nodes and 459 edges which roughly clustered into two sets depending on the types of genes were up-regulated or down-regulated.
The log2FC $> 1$ and p-value $< 0.01$ as significant and all 5 genes met the statistical standard. We compared cancer and normal bladder in GEPIA database using data from TCGA and GTEx. We considered EXPRESSION of CDK1, UBE2C, CDCA8 between bladder cancer and normal bladder in GEPIA database using data from TCGA and GTEx. We compared the mRNA expression of CDCA8, CDCA20, TPX2 and CDCA8 between bladder cancer and normal bladder in GEPIA database using data from TCGA and GTEx.
Comparison of CDK1 Across 11 Analyses

| Median Rank | p-Value | Gene |
|-------------|---------|------|
| 379.0       | 0.015   | CDK1 |

Comparison of UBE2C Across 11 Analyses

| Median Rank | p-Value | Gene |
|-------------|---------|------|
| 224.0       | 9.23E-7 | UBE2C|

Comparison of CDC20 Across 11 Analyses

| Median Rank | p-Value | Gene |
|-------------|---------|------|
| 199.0       | 2.07E-8 | CDC20|

Comparison of TPX2 Across 11 Analyses

| Median Rank | p-Value | Gene |
|-------------|---------|------|
| 2083.0      | 0.171   | TPX2 |

Comparison of CDC48 Across 11 Analyses

| Median Rank | p-Value | Gene |
|-------------|---------|------|
| 696.0       | 1.22E-13| CDC48|

Legend

1. Infiltrating Bladder Urothelial Carcinoma vs. Normal Bladder, Clin Cancer Res, 2005
2. Superficial Bladder Cancer vs. Normal Bladder, Clin Cancer Res, 2005
3. Infiltrating Bladder Urothelial Carcinoma vs. Normal Dysplastic Bladder, Cancer Res, 2004
4. Stage 0 Bladder Urothelial Carcinoma vs. Normal Dysplastic Bladder, Cancer Res, 2004
5. Superficial Bladder Cancer vs. Normal Dysplastic Bladder, Cancer Res, 2004
6. Infiltrating Bladder Urothelial Carcinoma vs. Normal Lee Bladder, J Clin Oncol, 2010
7. Superficial Bladder Cancer vs. Normal Lee Bladder, J Clin Oncol, 2010
8. Infiltrating Bladder Urothelial Carcinoma vs. Normal Modlich Bladder, Clin Cancer Res, 2004
9. Superficial Bladder Cancer vs. Normal Modlich Bladder, Clin Cancer Res, 2004
10. Infiltrating Bladder Urothelial Carcinoma vs. Normal Sanchez-Carbajo Bladder, J Clin Oncol, 2006
11. Superficial Bladder Cancer vs. Normal Sanchez-Carbajo Bladder, J Clin Oncol, 2006

The rank for a gene is the median rank for that gene across each of the analyses.

The p-Value for a gene is its p-Value for the median-ranked analysis.
(a) High antibody staining immunohistochemistry images of 4 hub genes are available at https://www.proteinatlas.org/ENSG00000170312-CDK1/pathology/urothelial+cancer#img (CDK1), https://www.proteinatlas.org/ENSG00000175063-UBE2C/pathology/urothelial+cancer#img (UBE2C), https://www.proteinatlas.org/ENSG00000117399-CDC20/pathology/urothelial+cancer#img (CDC20) and https://www.proteinatlas.org/ENSG00000134690-CDCA8/pathology/urothelial+cancer#img (CDCA8) in The Human Protein Atlas, respectively.

(b) 2 antibodies were used for CDK1 (HPA003387 and CAB003799) and CDC20 (CAE004525 and HPA055288) and 3 for UBE2C (CAB011464, CAB035990 and HPA054975) and CDCA8 (HPA028120, HPA028258 and HPA028783). High positive staining rate was detected in urothelial cancer tissues of bladder for CDK1 (44/46, UBE2C (57/57), CDC20 (43/47), and CDCA8 (57/61), however, only CDC20 (P=0.006) and CDCA8 (P=0.007) showed significant results by comparing normal urinary bladder with urothelial cancer tissues of bladder using Mann-Whitney Test. The cutoff P value was set as 0.05.
Legends

We conducted survival analysis using GEO data with R language software and TCGA data in Oncolnc database to demonstrate the association between the expression of significant hub genes and overall survival time. CDC20 showed a statistical significance in both database (P=0.001 in GEO; P=0.0361 in Oncolnc), while CDCA8 only showed a statistical significance in GEO (P<0.001 in GEO; P=0.277 in Oncolnc).
Table 1 Characteristics of included datasets

| Dataset   | Platform                  | Bladder caner | Normal |
|-----------|---------------------------|---------------|--------|
| GSE7476   | GPL570                    | 9             | 3      |
|           | Affymetrix Human Genome U133 Plus 2.0 Array |               |        |
| GSE13507  | GPL6102                   | 188           | 67     |
|           | Illumina human-6 v2.0 expression beadchip |               |        |
| GSE40355  | GPL13497                  | 16            | 8      |
|           | Agilent-026652 Whole Human Genome Microarray 4x44K v2 |               |        |
| GSE52519  | GPL6884                   | 9             | 3      |
|           | Illumina HumanWG-6 v3.0 expression beadchip |               |        |
| GSE65635  | GPL14951                  |               |        |
|           | Illumina HumanHT-12 WG-DASL V4.0 R2 expression beadchip | 8             | 4      |
## Table 2 Top five hub genes with the highest degree of connectivity

| Rank | Gene symbol | Gene description                          | Connectivity degree |
|------|-------------|-------------------------------------------|---------------------|
| 1    | CDK1        | Cyclin-Dependent Kinase 1                 | 30                  |
| 1    | UBE2C       | Ubiquitin-Conjugating Enzyme E2 C         | 30                  |
| 3    | CDC20       | Cell Division Cycle 20                    | 29                  |
| 3    | TPX2        | Microtubule Nucleation Factor             | 29                  |
| 3    | CDCA8       | Cell Division Cycle Associated 8          | 29                  |
Gene expression value

Sample
- case
- control

Samples

GSE65635
