Identification of essential genes and drug discovery in bladder cancer and inflammatory bowel disease via text mining and bioinformatics analysis

Qingyuan Zheng
Department of Urology, Renmin Hospital of Wuhan University

Liantao Guo
Department of Breast and Thyroid Surgery, Renmin Hospital of Wuhan University

Rui Yang
Department of Urology, Renmin Hospital of Wuhan University

Xinmiao Ni
Department of Urology, Renmin Hospital of Wuhan University

Song Yang
Department of Urology, Renmin Hospital of Wuhan University

Lei Wang
Department of Urology, Renmin Hospital of Wuhan University

Zhiyuan Chen
Department of Urology, Renmin Hospital of Wuhan University

Xiuheng Liu (✉ drliuxh@hotmail.com)
Department of Urology, Renmin Hospital of Wuhan University

Research Article

Keywords: bladder cancer, inflammatory bowel disease, crohn's disease, text mining, drug discovery

Posted Date: July 1st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1777444/v1

License: © This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Abstract

Background: Bladder cancer (BCa) is the most common malignancy of the urinary system. Inflammation is critical in the occurrence and development of BCa. The purpose of this study was to identify key genes and pathways of inflammatory bowel disease in BCa through text mining technology and bioinformatics technology, and to explore potential therapeutic drugs for BCa.

Methods: Genes associated with BCa and Crohn's disease (CD) were detected using the text mining tool GenClip3, and analyzed using Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG). A protein-protein interaction (PPI) network was constructed by STRING and visualized in Cytoscape, and modular analysis was performed using the Molecular Complex Detection plugin (MCODE). Finally, the genes clustered in the first two modules were selected as core genes, and the drug-gene interaction database was used to discover potential therapeutic drugs.

Results: We identified 796 genes shared by "Bladder cancer" and "Crohn's disease" by text mining. Gene function enrichment analysis yielded 18 enriched GO terms and the 6 most relevant KEGG pathways. A PPI network with 758 nodes and 4014 edges was constructed, and 20 gene modules were obtained using MCODE. We selected the top two gene clusters as core candidate genes. We found that 3 out of 55 selected core genes could be targeted by 26 existing drugs.

Conclusions: The results indicated that CXCL12, FGF2 and FSCN1 are potential key genes involved in CD with BCa. Additionally, 26 drugs were identified as potential therapeutics for BCa treatment and management.

Introduction

Bladder cancer (BCa) is the most common tumor of the urinary system and one of the top ten cancers worldwide, with approximately 573,000 new cases and 213,000 deaths in 2020.[1] Several factors are known to influence BCa risk, including smoking, occupational exposure to carcinogenic aromatic amines, high triglycerides, and schistosomiasis.[2] In addition, studies have shown that inflammatory bowel disease (IBD), especially Crohn's disease (CD), may increase the overall risk of BCa.[3, 4] At present, chemotherapy plays a crucial role before or after surgery, which can help control lesions before surgery to facilitate surgery, and relieve clinical symptoms after surgery.[5–7] However, many chemotherapy drugs (BCG, gemcitabine and mitomycin, etc.) have great side effects, can produce significant systemic toxicity and local toxicity, often lead to patients intolerable.[8] Immunosuppressive drugs, which are widely used in the treatment of IBD, have been reported as contributing to the increased risk of BCa and other extraintestinal cancers.[9–11]

IBD is a heterogeneous group of immune-mediated diseases. It can be divided into CD and ulcerative colitis (UC) based on clinical features, histological and pathological features, and laboratory and endoscopic findings.[12] Smoking is a major risk factor for CD and can lead to reduced response to drug therapy.[13] Studies have demonstrated a link between IBD and extraintestinal cancers observed in IBD.
patients.[14] Inflammation is a common host response to malignant bladder tumors, yet it also promotes BCa proliferation, invasion, migration, and angiogenesis, and leads to chemoresistance.[15] Therefore, effective CD treatment may be a potential therapy option for BCa prevention.

Computerized text mining technology refers to the use of natural language processing, artificial intelligence, information retrieval and data mining methods to discover, retrieve and extract information from a text corpus. It is an effective tool for comprehensive analysis and interpretation of literature and high-throughput data in the era of big data.[16–18] The maturing and gradually popularized text mining technology makes it possible to automatically collect disease-gene associations from a large number of literatures related to BCa and IBD, which is helpful to explore the therapeutic drugs and pharmacological targets of bladder cancer.

In this study, we explored the essential genes related to BCa and CD based on text mining technology and bioinformatics methods, and screened out the drugs for BCa, providing ideas for BCa treatment and new drug development and reference. First, we used text mining techniques to obtain common genes between BCa and CD. The in-depth information of these genes was integrated with pathway enrichment and functional analysis. Protein and protein interaction (PPI) analysis was then performed to identify important modular genes with more interactions. Finally, drug-gene interactions of important module genes were explored in the Drug-Gene Interaction Database (DGIdb) to find the candidate drugs we needed. Figure 1 shows the flow chart of this study.

**Methods**

**Computerized text mining**

Text mining was performed using GenClip3 (http://cismu.net/genclip3/), which was a publicly available website. GenClip3 links more than 4,000,000 articles in PubMed to nearly 160,000 genes in humans, making it easier for researchers to find literature that intersects a specific genome.[19] In the GenCLiP3 database, we took "All human genes" as the gene dataset, selected "Search in MEDLINE", and searched for the keywords "Bladder cancer" and "Crohn's disease" respectively. After retrieving the relevant genes, Venny (https://bioinfogp.cnb.csic.es/tools/venny/) was used to get the gene set that intersected the two, and saved this text mining genes (TMGs) for further data analysis.

**Gene ontology and pathway enrichment analysis**

Gene ontology (GO) analysis is a general and effective method to annotate gene products and their functional characteristics. GO analysis can be divided into biological process (BP), cellular component (CC) and molecular function (MF) according to different functions. The Kyoto Encyclopedia of Genes and Genomes (KEGG) is another publicly available database resource that integrates genomic, chemical, and systemic functional information.[20] In this study, we imported TMGs into METASCAPE (https://metascape.org) for GO and KEGG pathway annotation and enrichment analysis, and used for gene annotation visualization.
Protein-protein interaction network and module analysis

The interaction search tool STRING (https://string-db.org/cgi/input.pl) is a public database designed to integrate all known and predicted associations between proteins, including physical interactions and functional associations.[21] We input the TMGs into the STRING database (version 11.0), set the species as "Homo sapiens", and set the confidence level to 0.9 for screening to obtain PPI information of TMGs. The PPI information was then imported into Cytoscape software, and the important gene modules were clustered using molecular complex detection (MCODE) and STRING applications of Cytoscape software. Finally, we obtained modules that were densely connected in the PPI network, which were core genes associated with BCa and CD.

Gene data analysis

Gene Expression Profile Interaction Analysis (GEPIA, http://gepia.cancer-pku.cn/index.html) is a web-based tool that enables comprehensive expression analysis based on RNA-sequencing data from TCGA and GTEx projects.[22] We performed differential expression analysis and patient survival analysis on the core genes of PPI species by GEPIA, and obtained genes with significant differences in expression in both analyses for further study.

Drug-gene interactions

We entered the screened genes with significant differences into the Drug-Gene Interaction Database (DGIdb, https://www.dgidb.org/). DGIdb is an open and free web resource that includes different data sources describing drug-gene interactions and gene druggability to explore drugs or treatments that may have potential targets. Therefore, with the help of DGIdb, we could screen for possible new treatment options for BCa patients with CD.

Results

Identification of TMGs

Using the GenClip3 strategy described in the Methods section, we obtained 3042 unique genes associated with BCa and 1323 genes associated with CD; among them, 796 genes were associated with both diseases, namely TMGs.

GO and KEGG enrichment analysis of TMGs

GO bioprocess and KEGG pathway enrichment analyses were performed on TMGs using the METASCAPE website to identify the most enriched terms most closely associated with BCa and CD. Figure 2 shows the top six significant enrichment terms for BP, CC and MF. The six most enriched terms were ‘positive regulation of cytokine production’, ‘positive regulation of protein phosphorylation’, ‘inflammatory response’, ‘regulation of defense response’, ‘response to lipopolysaccharide’, ‘cellular response to cytokine stimulus’. As for the CC annotation, it was significantly enriched on ‘external side of plasma membrane’,
‘extracellular matrix’, ‘membrane raft’, ‘vesicle lumen’, ‘endocytic vesicle’, ‘receptor complex’. In the MF category, the following items were most enriched: ‘signaling receptor activator activity’, ‘kinase binding’, ‘transcription factor binding’, ‘protein homodimerization activity’, ‘cytokine binding’, ‘G protein-coupled receptor binding’. Subsequently, we analyzed the enriched KEGG signaling pathway annotations and selected 6 significantly enriched pathways for presentation in Fig. 2. The six items were ‘Pathways in cancer’, ‘Cytokine-cytokine receptor interaction’, ‘Lipid and atherosclerosis’, ‘PI3K-Akt signaling pathway’, ‘IL-17 signaling pathway’, ‘AGE-RAGE signaling pathway in diabetic complications’.

**PPI network construction and modular analysis**

Protein-protein interaction analysis was performed using the STRING website to construct a PPI network for TMGs. From the TMGs obtained above as input, STRING visualizes their interactions with confidence scores set to > 0.900. The network has a total of 758 nodes and 4014 edges (Supplementary Fig. 1). The network was then sent to Cytoscape for modular analysis using the built-in MCODE program, and a total of 20 clusters were obtained. We selected the first-ranked Cluster 1 (21 nodes and 138 edges, score 13.8) and the second-ranked Cluster 2 (34 nodes and 136 edges, score 8.24) for subsequent identification of core genes (Fig. 3).

**Gene data analysis**

Cluster 1 and Cluster 2 contained a total of 55 genes. We input 55 genes into GEPIA for gene expression differential analysis and patient survival analysis, and the results were presented in the form of boxplots and Kaplan-Meier curves. The expressions of IL6ST, IL6R, IL1A, CXCL10, CSF3, CCL2, BCL6, BIRC5, CXCL12, FGF2, FSCN1, FOXO1, HGF, JAK2, MAP1LC3A, MMP1, MMP2, MMP9, STAT5B, VIM were significantly different between the normal group and the tumor group, and the results of the survival analysis showed that only the expressions of CXCL12, FGF2 and FSCN1 were significantly correlated with the overall survival of BCa patients (Fig. 4, Supplementary Figure S3–S11). Therefore, we could select CXCL12, FGF2 and FSCN1 as the core genes to be screened for drug-gene interaction analysis.

**Discussion**

We identified three core genes as potential targets for drug-gene interaction analysis and obtained an initial list of 26 drugs, including anticancer drugs, anti-inflammatory drugs, antiulcer agents, antihistamines, etc. (Fig. 5, Table 1). Seven drugs were associated with CXCL12, 19 drugs were associated with FGF2, one drug was associated with FSCN1, and one drug was associated with both CXCL12 and FGF2 genes. Among these candidate drugs, some approved anticancer compounds have been screened, such as THALIDOMIDE and VINCristine; some anti-ulcer drugs, such as SUCRAlFATE, which is a gastric mucosal protective drug, have the ability to protect the ulcer surface and promote ulcer healing. [23]
BCa is a disease with complex molecular features and high morbidity and mortality. As one of the factors influencing the increased risk of BCa, CD has been shown to promote the progression of BCa and other extraintestinal cancers. In this study, we used text mining technology and bioinformatics technology to identify three key genes related to BCa and CD, namely CXCL12, FGF2 and FSCN1, and searched for 26 drugs that may have potential therapeutic effects.

Currently, a growing body of research suggests that many human cancers are caused by chronic infection or a persistent state of chronic inflammation. Inflammatory factor storm can stimulate the secretion of tumor growth-promoting factors while attenuating the cytotoxic function of immune cells. As an immune-mediated persistent chronic inflammatory disease, IBD can also increase the risk of intestinal or extraintestinal tumors. Therefore, selection of appropriate candidate drugs is critical to maximize the clinical benefit of inflammation-based therapeutic strategies.

The CXCL12 gene is a chemokine, also known as stromal cell-derived factor-1, involved in various physiological and pathological processes, including embryogenesis, neurogenesis, hematopoiesis, angiogenesis, lympho-genesis, and inflammation. Experiment-based studies had shown that enhanced expression of CXCL12 was closely related to the development, invasion and metastasis of BCa. These current studies suggested that CXCL12 might be a prognostic marker and potential immunotherapy target for BCa. Fibroblast growth factor 2 (FGF2) is an important pro-angiogenic factor that has been found to be expressed in a variety of human cancers and promote angiogenesis by interacting with receptors expressed on the surface of endothelial cells. Elevated FGF2 expression was reported to be associated with aggressive pathological features and poorer outcomes after radical cystectomy, and was an independent prognostic factor for disease recurrence. Fascin actin-bundling protein 1 (FSCN1) plays an important role in the regulation of cell migration and invasion during BCa progression. More and more studies had shown that inhibition of FSCN1 might serve as a promising therapeutic target for invasion and metastasis of BCa. In a word, our study provides the basis for potential therapeutic strategies targeting key genes CXCL12, FGF2 and FSCN1.

According to our study, the screened 26 drugs may act in multiple ways to achieve potential therapeutic effects on IBD in BCa patients. Aspirin is a non-steroidal anti-inflammatory drug (NSAIDs) that has been reported to reduce the risk of BCa when taken regularly. BCa frequently overexpresses COX-2, and NSAIDs are potent inhibitors of COX enzymes, which can inhibit apoptosis and stimulate the NF-κB pathway of COX-2, and can also induce peroxisome proliferators Activating receptors suppress tumors. Vincristine, an alkaloid screened from medicinal plants and herbs, exhibits antiproliferative and antitumor effects. Studies had shown that neoadjuvant chemotherapy (CMV) with vincristine combined with cisplatin and methotrexate significantly prolonged survival in patients with invasive BCa. Another study showed that the same CMV chemotherapy regimen significantly reduced the risk of death in patients with invasive bladder cancer. Therefore, CMV chemotherapy as first-line adjuvant therapy for invasive bladder cancer may improve prognosis. In short, the drugs screened by us may provide new ideas for future treatments.
There are some limitations of our study. First, our experiments are based on some publicly available databases, such as TCGA database and GTEx database, and larger-scale data are needed to further confirm our research. Second, we only used computerized text mining technique and bioinformatics technique to study and search for potential therapeutic drugs and targets of BCa, and further in vivo and in vitro experiments are needed to demonstrate the key roles of these genes and drugs in BCa.

In conclusion, we identified CXCL12, FGF2 and FSCN1 as key candidate genes related to BCa and CD, and screened out 26 drugs targeting candidate genes through text mining and analysis of different databases. The rational use of candidate drugs and the combination of existing drugs may bring great benefits to the treatment of BCa.

Declarations

Acknowledgements

The authors wish to thank all the researchers and colleagues who provided the data.

Conflicts of interest

There are no conflicts of interest.

Author contributions

XHL reports being Director and the sponsor of this study. QYZ, LTG and RY performed the research, analysed the data and wrote the paper. XMN, SY, LW and ZYC contributed essential reagents or tools.

Availability of data and material

All materials and data may be obtained from the corresponding author upon reasonable request.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

References


42. Wahli W, Michalik L. PPARs at the crossroads of lipid signaling and inflammation. Trends Endocrinol Metab. 2012;23(7):351-363.10.1016/j.tem.2012.05.001
Tables

Table 1 is available in the Supplementary Files section.

Figures
Text mining technology was used to obtain the common genes of bladder cancer and crohn's disease, namely text mining genes. These genes were then integrated by GO functional analysis and KEGG enrichment. PPI analysis was performed to identify important modular genes. Final candidate drugs were
performed in DGIdb by gene-drug interaction analysis. KEGG, Kyoto Encyclopedia of Genes and Genomes; GO, Gene Ontology; PPI, Protein-protein interaction; DGIdb, Drug-Gene Interaction Database.

Figure 2

Results of GO and KEGG enrichment analysis.

Top six significant enrichment terms for BP, CC, MF and KEGG signaling pathways. The size of the circles corresponds to the gene counts enriched in each term. KEGG, Kyoto Encyclopedia of Genes and Genomes; GO, Gene Ontology.
Figure 3

The top two highest scoring gene modules (Cluster 1 and Cluster 2) were selected using Cytoscape. Cluster 1 contains 21 nodes and 138 edges generated by the MCODE application, and cluster 2 contains 34 nodes and 136 edges.
Figure 4

Expression and overall survival of CXCL12, FGF2 and FSCN1. (a) The expressions of CXCL12, FGF2 and FSCN1 were significantly different between normal and tumor groups. (b) High and low expression groups of CXCL12, FGF2, and FSCN1 in patients were significantly associated with poorer overall survival (P < 0.005).
Figure 5

26 drugs with potential therapeutic effects on Crohn's disease in bladder cancer patients were collected. Using the final list of two genes as potential targets in the drug-gene interaction analysis, a list of drugs was selected as possible drug treatments.

**Supplementary Files**

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

- Table1.xlsx
- Supplementarymaterial.pdf