Development and validation of a novel stem cell subtype for bladder cancer based on stem genomic profiling

Chaozhi Tang  
China Medical University Hospital

Jiakang Ma  
Zhengzhou University

Xiuli Liu  
Guilin Medical University Affiliated Hospital

Zhengchun Liu (✉️ 18624879717@163.com)  
Guilin Medical University Affiliated Hospital

Research

Keywords: bladder cancer, stem cells, immune microenvironment, hypoxia, epithelial mesenchymal transition, immunotherapy

Posted Date: May 18th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-28647/v1

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

Version of Record: A version of this preprint was published at Stem Cell Research & Therapy on October 28th, 2020. See the published version at https://doi.org/10.1186/s13287-020-01973-4.
Abstract

Background Bladder cancer is the fifth most common type of cancer worldwide, with high recurrence and progression rates. Although considerable progress has been made in the treatment of bladder cancer through accurate typing of molecular characteristics, little is known about the various genetic and epigenetic changes that have evolved in stem and progenitor cells. Thus, we developed a novel stem cell typing method to fill this gap. Methods Based on six published genomic data sets, we used 26 stem cell gene sets to classify each data set. Unsupervised and supervised machine learning methods were used to perform the classification. Results We classified BLCA into three subtypes—high stem cell enrichment (SCE_H), medium stem cell enrichment (SCE_M), and low stem cell enrichment (SCE_L)—based on multiple cross-platform data sets. The stability and reliability of the classification were verified. The stemness index obtained from the one-class logistic regression machine learning method showed that the degree of tumor stem cell enrichment was not proportional to the stemness index. Compared with the other subtypes, SCE_H showed the highest degree of cancer stem cell concentration, lowest stemness index, and highest level of immune cell infiltration and was the most sensitive to predicted immune checkpoint inhibitor treatment. However, this group showed the worst prognosis. Comparison of gene set enrichment analysis results for pathway enrichment of various subtypes revealed that the SCE_H subtype activates some important pathways regulating cancer occurrence, development, and even poor prognosis, including epithelial mesenchymal transition, hypoxia, angiogenesis, KRAS signal up-regulation, the interleukin 6-mediated Jak-STAT signaling pathway, and inflammatory response. Two identified pairs of transcription factors, GRHL2 and GATA6 and IRF5 and GATA3, likely have opposite regulatory effects on SCE_H and SCE_L, respectively. Conclusions The identification of BLCA subtypes based on cancer stem cell gene sets revealed the complex mechanism of carcinogenesis causing BLCA and provides a new direction for the diagnosis and treatment of BLCA.

Background

Bladder cancer (BLCA) generally occurs in bladder intraepithelial cells and is the fifth most common type of cancer worldwide. It is estimated that there are 151,000 new cases of BLCA and more than 52,000 related deaths each year [1–3]. Urothelial cancer is the most common type of BLCA, accounting for approximately 90% of all BLCA cases [1]. Most patients with BLCA can be diagnosed early, but the rate of recurrence and progression is high. Approximately 78% of patients relapse within 5 years [4]. With the rapid development of molecular biology and continuous emergence of biological detection technology, molecular typing of BLCA through genetic analysis has revealed differences in drug reactivity and the prognosis of patients with BLCA based on their biological heterogeneity, which has also made some long-term progress. For example, molecular classification of the Cancer Genome Atlas Quartile [5], University of North Carolina Dichotomy [6], MD Anderson Cancer Center Trisection [7], and Lund University Quintiles [8]. Although these classification methods reveal the pathogenic mechanism of BLCA at the molecular level, they do not fundamentally demonstrate the origin of heterogeneity in tumors. Emerging evidence suggests that cancer stem cell (CSC) subpopulations are characterized by a mixture of stem cells and
cancer cells. In addition to the abilities of self-renewal and differentiation, CSCs can also act as seeds of tumors [9, 10], and thus are considered as the driving force of heterogeneity.

The tumor is a complex integrated system composed of relatively differentiated tumor cells, infiltrating immune cells, CSCs, tumor-associated endothelial cells, stromal cells, and other cell types [11, 12]. The function and plasticity of CSCs may be induced by specific microenvironment signals and cell interactions in the tumor niche. Studies have shown that stem cells in melanoma can preferentially inhibit T cell activation and support the induction of regulatory T cells, thereby evading recognition by the immune system [13]. In glioblastoma, CSCs suppress T cell responses by generating immunosuppressive cytokines through the STAT3 pathway and inducing apoptosis of T cells, leading to increased cancer stemness and carcinogenic potential [14]. Additionally, some molecular signal transduction pathways that control the normal balance of normal stem cells are abnormally activated or inhibited to contribute to the self-renewal, proliferation, survival, and differentiation characteristics of CSCs. In a study of an experimental model of colon cancer, elevated inflammatory nuclear factor κB signal transduction enhanced Wnt activation and induced dedifferentiation of non-stem cells, which acquired tumor-initiating ability [15]. These results indicate that immune cells and their related cytokines and signal transduction pathways can directly regulate and enhance the CSC phenotype.

In our study, we divided BLCA into high stem cell enrichment (SCE_H), medium stem cell enrichment (SCE_M), and low stem cell enrichment (SCE_L) subtypes, using stem cell gene sets collection. We demonstrated the stability and reliability of this classification using six independent data sets using the unsupervised clustering method. Importantly, we systematically examined the prognostic significance of BLCA stem cell subtypes, relationship between immune cells and genes, sensitivity of immune checkpoint inhibitor treatment, and possible changes in biological pathways and important transcriptional regulation factor/network (Fig. 1). Our results provide insight into and a basis for studies of BLCA stem cells as well as useful information for the clinical diagnosis and treatment of BLCA.

**Methods**

**Stem cell signature collection**

The 26 human stem cell gene sets used in this study were obtained from StemChecker (http://stemchecker.sysbiolab.eu/) [16] Expression Checks (18), RNAi Screens (1), Literal Curation (2), Computationally Derived (2), and TF Target Genes (3).

**Data processing**

The data sets used to identify the BLCA stem cell subtypes were from three different platforms: The Cancer Genome Atlas (TCGA), Gene Expression Omnibus (GEO), and ArrayExpress databases. TCGA’s RNA-seq data (fragments per kilobase of transcript per million mapped reads, FPKM) of 19 normal samples and 414 cancer samples and clinical information were downloaded from TCGA Knowledge Base (https://portal.gdc.cancer.gov/repository). Gene annotation was performed using the Ensemble
database. The ArrayExpress database contains RNA-seq and clinical data (N = 476) for 476 cases of early urothelial carcinoma (E-MTAB-4321) FPKM from the European Genome-phenome Archive. The expression matrices of four GEO datasets, GSE13507 (N = 256), GSE32548 (N = 131), GSE31684 (N = 93), and GSE32894 (N = 308), were all quantile-normalized, and the genes were annotated in their respective platform files Illumina human -6 v2.0 expression beadchip, Illumina HumanHT-12 V3.0 expression beadchip, [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array, and Illumina HumanHT-12 V3.0 expression beadchip.

**Identification of BLCA subtypes based on stem cell gene sets**

For each BLCA data set, we used the GSVA package to perform single-sample gene set enrichment analysis (ssGSEA) to quantify the enrichment level of each BLCA sample in the 26 stem cell gene sets. The ConsensusClusterPlus package was used for consensus clustering and stem cell subtype screening of the ssGSEA scores. Briefly, k-means clustering was performed with 50 iterations (each using 80% of samples). The best cluster number was determined by the clustering score for the cumulative distribution function (CDF) curve, and the relative changes in the area under the CDF curve were evaluated.

**Survival analysis**

The Kaplan–Meier curve is used to describe the difference in survival time of patients with BLCA in different data sets for classifying stem cell subtypes. We compared the survival prognosis of patients with BLCA (overall survival (OS), relapse-free survival (RFS), and progression-free survival (PFS)). The log-rank test used $P < 0.05$ as the threshold to detect significant differences in survival time.

**Immune checkpoint inhibitor treatment response prediction**

Tumor immune dysfunction and exclusion is a calculation method for simulating tumor immune escape mainly by testing how the expression of each gene in the tumor interacts with the level of cytotoxic T cell (CTL) infiltration to affect patient survival [17]. We used TCGA's FPKM RNA_seq expression profile combination subclass mapping method to predict the clinical response of BLCA stem cell subtypes to immune checkpoint blockade [18].

**Calculation of stemness index of BLCA based on gene expression profile**

Malta TM [19] uses the OCLR algorithm to establish a predictive model for pluripotent stem cell samples to train expression profile stemness (mRNAsi) based on a gene expression profile containing 11,774 genes from human normal tissues and cancer tissues and introduces the workflow (https://bioinformaticsfmrp.github.io/PanCanStem_Web/) for obtaining the stemness index. We used the same Spearman correlation operator for each BLCA data set and applied the stemness index model to score the BLCA samples. Using the stemness index, we verified the relationship between stem cell subtypes and clinical traits.

**Pathway enrichment analysis**
We compared the biological changes in every two subtypes in TCGA data set, and used h.all.v7.1.symbols.gmt as the reference gene set for gene set enrichment analysis (GSEA). Analysis was performed with 1000 permutations, a <0.05 false discovery rate as the screening threshold, and GSEA version 4.0.1.

**Evaluation of immune cell infiltration level, tumor purity, and stromal content in BLCA**

ESTIMATE was used to evaluate the level of immune cell infiltration, tumor purity, and stromal content in the BLCA stem cell typing [20].

**Comparison of immune cell fraction between BLCA stem cell subtypes**

CIBERSORT is an algorithm that deconvolves the expression matrix of 22 human immune cell subgroups and can be used to estimate the proportion of immune cells [21]. We set the permutations to 1000 and used \( P < 0.05 \) as the screening threshold. The Kruskal-Wallis test was used to compare the differences in immune cell components before each subtype of BLCA stem cell subtypes.

**Gene co-expression network analysis**

To identify key genes or gene networks that are characteristic of various stem cell subtypes in BLCA, we performed weighted correlation network analysis (WGCNA) [22] to detect gene modules related to stem cell subtypes. The gene matrix is composed of 4876 differential genes in BLCA control normal tissues (difference is generated by limma package in R, \(|\log_2 \text{fold-change}| > 1, P < 0.05\)). WGCNA network construction and module detection used the unsigned topological overlap matrix, the best soft threshold (power) was selected as 3, the minimum number of genes in the module was 50, and the branch merge interception height was 0.25. The hub gene is defined as that which has a Pearson correlation (in practice, because of the generally low value of the connection weight, the Pearson correlation is used) of greater than 0.30, and at least 10 genes connected to it. The gene co-expression network was visualized using Cytoscape 3.7.1 software. Kruskal–Wallis tests were used to examine the expression differences of hub genes between stem cell subtypes. The results of survival analysis were divided into high and low groups based on median expression of transcription factor using GEPIA (http://gepia.cancer-pku.cn/) database [23]. Log rank test was used. Top 20 enrichment pathways obtained by using Metascape (http://metascape.org/gp/index.html#/main/step1).

**Statistical analysis**

Comparison of the stemness index of BLCA stem cell subtypes was performed using the Kruskal-Wallis test, comparison of the cancer and normal tissue stemness index was performed using the unpaired \( t \) test, and comparison of the BLCA molecular subtype and stemness index of T stage was performed using analysis of variance (ANOVA) in GraphPad Prism 8.0. \( CD274 \) expression differences between stem cell subtypes was evaluated by ANOVA in R. All tests were two-tailed, and \( P < 0.05 \) was considered as statistically significant.
Results

BLCA subtypes identified based on stem cell gene sets

We collected 26 stem cell gene sets representing unique self-regenerating properties (Supplementary Table 1) and quantified the scores of 26 stem cell gene sets in each sample by ssGSEA. We used the ConsensusClusterPlus package to divide all tumor samples into k (k = 2–9) different subtypes. The CDF curve based on the consensus scores achieves the best division when k = 3. Additionally, the principal component analysis results indicated that the ssGSEA scores based on the 26 stem cell gene sets were clearly divided into 3 subtypes (Figure 2A–D). We defined these three subtypes as SCE_H, SCE_M, and SCE_L (Figure 2E). Similarly, we performed the same clustering and subtyping for the remaining data sets E-MATB-4321, GSE13507, GSE31684, GSE32548, and GSE32894.

When using ESTIMATE to evaluate the level of immune infiltration in all data sets, we found that the SCE_H immune score in all 6 data sets was much higher than that for other subtypes, and SCE_L showed the lowest immune score (Supplementary Figure 1). Comparison of the stromal content revealed the same trend (SCE_H > SCE_M > SCE_L). However, comparison of tumor purity of the BLCA stem cell subtypes showed the opposite results. SCE_H and SCE_L showed the lowest and highest tumor purity (SCE_L > SCE_M > SCE_H), respectively. This is consistent with the results observed for most HLA genes and immune cell marker genes evaluated, such as CD8A (CD8 T cells), GZMA (cytotoxic cells), IFNG (Th1 cells) PMCH (Th2 cells), CD68 (macrophages), and IL17A (Th17 cells), among others, which were significantly up-regulated and down-regulated in SCE_H and SCE_L, respectively (Supplementary Figure 2).

Based on the close relationship between BLCA stem cell subtypes and immunity, we focused on the differential expression of CD274 (PD-L1) in each subtype (Figure 3). In all six data sets, SCE_H and SCE_L showed the highest and lowest expression levels, respectively, revealing that the BLCA subtype SCE_H was more sensitive to anti-PD-1 immunotherapy than the other subtypes. Subsequent immune checkpoints inhibitor treatment response prediction and survival analysis confirmed these results.

Survival of patients with different BLCA stem cell subtypes

Because BLCA is a heterogeneous disease with a high recurrence rate and shows disease progression, exploring the relationship between subtype classification and clinical prognosis is beneficial for the prognosis assessment and corresponding clinical management of BLCA. We performed OS, RFS, and PFS analysis of the six data sets. Unexpectedly, all data sets showed consistent trends (Figure 4). SCE_H and SCE_L showed the worst and best survival in prognostic analysis (SCE_L > SCE_M > SCE_H), respectively. The P values of the log rank for the OS of TCGA, GSE31684, GSE32548, and GSE32849 were 5.631e-4, 0.038, 2.158e-4, and 8.755e-6, respectively. The P-values for the log-rank of PFS for TCGA, E-MTAB-4321, and GSE13507 were 0.004, 3.976e-9, and 7.046e-4, respectively, and the log-rank P value of TCGA RFS was 0.032.
Prediction of therapeutic response of BLCA stem cell subtypes to immune checkpoint inhibitors

Based on the above results, we further evaluated the responses of the three subtypes to immunotherapy. At present, 5 PD-1/PD-L1 immunotherapy drugs have been approved by the Food and Drug Administration for treating BLCA, including nivolumab and pembrolizumab (both PD-1 inhibitors) approved in 2016 and 2017 for treating patients with locally advanced or metastatic urothelial cancer who were administered first-line platinum-containing chemotherapy for one year [24-29]. We used the tumor immune dysfunction and exclusion algorithm to predict the likelihood of a response to immunotherapy. The results showed obvious differences in the responses to immunotherapy among the SCE_H (20%, 32/158), SCE_M (42%, 71/168), and SCE_L groups (61%, 54/88) (P = 2.951e-10). We further performed subclass mapping to compare the expression profiles of the three stem cell subtypes which were defined using another published data set containing 47 patients with melanoma who responded to immunotherapy [30]. In pairwise comparison of the three subtypes, more promising results were observed in SCE_H for the anti-PD1 and anti-CTLA4 treatments compared to the other subtypes (Figure 5A–C) (anti-PD1 therapy: SCE_H vs SCE_L, FDR = 0.036; SCE_H vs SCE_M, p = 0.046; SCE_M vs SCE_L, FDR = 0.048; anti-CTLA4 therapy: SCE_H vs SCE_L, FDR = 0.036; SCE_H vs SCE_M, FDR = 0.008). We further correlated the BLCA stem cell typing results with the published molecular typing and immunotyping results in TCGA cohort. SCE_H mainly corresponded to the molecular subtypes luminal-infiltrated and basal squamous and C1 and C2 for immune subtypes; SCE_L mainly corresponded to luminal-papillary and C1–C4, and SCE_M showed a wide distribution of molecular and immune subtypes (Figure 5D).

Differences among 22 human immune cell subgroups of BLCA stem cell subtypes in CIBERSORT

To explore the reasons for the difference in survival of patients with different BLCA stem cell subtypes, we used the CIBERSORT algorithm to calculate the proportions of 22 immune cells in each subtype of the six data sets, with P <0 .05 as the threshold for screening. The results showed that the proportions of macrophages M0, M1, and M2 showed an upward trend in the SCE_H subtype (except for GSE13507) and the proportion of regulatory T cells (Tregs) was significantly (P <0 .05) increased in the SCE_H subtype (Figure 6A–F). We also used the ssGSEA scores of immune cells in each cohort as continuous variables and performed univariate Cox analysis (Supplementary Table 2). Further, we divided the median value of the ssGSEA scores of the corresponding data set into groups with high and low scores. A high score indicated that patients with a high macrophage M0 content had a worse prognosis. This is consistent with the poor clinical prognosis of patients with the SCE_H subtype compared to that of patients with the other subtypes. The trend for macrophages M2 was similar to that of macrophages M0, whereas macrophages M2 and Tregs showed opposite trends (Figure 6G–L). This reveals that compared to Tregs, macrophages M0 and M2 have completely opposite regulatory mechanisms in the prognosis of BLCA.

Correlation of CD274 with stemness genes and risk observation of stem cell subtype populations

We identified an important immune role for CD274 in the stem cell subtypes, and thus we further explored the correlation between CD274 and the identified stemness genes. Figures 7A–G show scatter plots of the expression of CD274 and stemness genes CD44, GATA3, HIF1A, ID1, MYC, SOX9, and CXCL8 in BLCA
TCGA cohort. Among them, \( CD274 \) was negatively correlated with \( ID1 \) and \( GATA3 \) and positively correlated with \( CD44, HIF1A, MYC, SOX9, \) and \( CXCL8 \) (Figure 7H). We divided the patients according to the optimal expression cut-off of \( CD274 \) and each stemness gene into risk groups I, II, III, and IV (for example, \( CD274 \) and \( CD44 \) corresponded to \( CD274^{\text{low}} CD44^{\text{low}}, CD274^{\text{high}} CD44^{\text{low}}, CD274^{\text{low}} CD44^{\text{high}}, \) and \( CD274^{\text{high}} CD44^{\text{high}} \)). According to the scatter plot, for each pair of risk groups divided by \( CD274 \) and the optimal threshold of stemness gene expression, patients with stem cell subtypes were mainly concentrated in groups I and III (87–88%), the SCE_L subtype was mainly concentrated in risk group III of the \( CD274 \) and \( GATA3 \) and \( ID1 \) pairs, and in the \( CD274 \) and \( CD44, HIF1A, MYC, SOX, \) and \( CXCL8 \) gene pairs were concentrated in risk group I. Further survival analysis of groups I and III of each gene pair showed that patients with the higher SCE_L subtype population had better survival than patients with a lower SCE_L subtype population (Figure 7I–O). This is consistent with observation that patients in the SCE_L group had the longest survival time.

**Stemness index based on mRNA expression (mRNAsi) in BLCA**

We evaluated the differences in stem cell subtypes defined based on the stemness index obtained from an existent one-class logistic regression (OCLR). Five of the six data sets (Figure 8A–E) showed the decreasing trend of SCE_M > SCE_L > SCE_H, which showed that different levels of stem cells enriched the potential for self-renewal and differentiation. In the E-MTAB-4321 data set with <T2 accounting for 96.6% (460 cases), the opposite results were observed. The stemness index decreased according to the trend of SCE_H > SCE_L > SCE_M (Figure 8F), reflecting the high impact of the T stage on the stemness index and revealing dynamic development of the tumor. This difference was also reflected in normal and cancer tissues, primary and recurrent (Figure 8G–I); thus, the stemness index of cancer tissue is much higher than that of normal tissue, with recurring cancer showing a higher index than primary cancer. Even in the non-uniform molecular classification of BLCA, the stemness index can effectively characterize the characteristics of different subtypes. According to the BLCA subtypes described by Robertson, neuronal and luminal infiltrated subtypes have the highest and lowest stemness indices respectively, which is consistent with corresponding reports showing that the neuronal type has the worst prognosis compared to other types and a better prognosis for the luminal infiltrated type. Similarly, among the five BLCA subtypes classified by Lund University [8], the genomically unstable and infiltrated types have the highest and lowest stemness indices, respectively (Figure 8J–K). Figures 8L–P show the difference in the T stage stemness index in each data set.

**GSEA for BLCA stem cell subtypes**

To explore the biological changes caused by differences in the enrichment of stem cells, we conducted pairwise comparison of the GSEA results for each subtype.

By selecting at least one pathway with an FDR <0.05 for the same pathway, we found that as the enrichment of stem cells increased, epithelial mesenchymal transition (EMT) became more significant. EMT is considered as a signal of malignant transformation in all cancers, giving cells the ability to
metastasize and invade, including by imparting stem cell characteristics, reducing apoptosis and aging, and resisting chemical and immunotherapy [31, 32] (Table 1). EMT can also activate multiple pathways, regulate cell metabolism, angiogenesis, proliferation, and migration, and enable cells to respond hypoxic environments. Pathways are also significantly enriched, for example, in hypoxia, angiogenesis, inflammatory response, IL6-mediated Jak-STAT signaling pathway, and KRAS signal up-regulation (Figure 9). These pathways together constitute a vicious circle of cancer occurrence, proliferation, invasion, and metastasis.

Table 1. GSEA for BLCA stem cell subtypes.
<table>
<thead>
<tr>
<th>pathway</th>
<th>NES</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCE_H vs SCE_L</td>
<td>2.21</td>
<td>0.005</td>
</tr>
<tr>
<td>SCE_H vs SCE_M</td>
<td>2.42</td>
<td>0</td>
</tr>
<tr>
<td>SCE_M vs SCE_L</td>
<td>2</td>
<td>0.008</td>
</tr>
<tr>
<td><strong>HALLMARK_HYPOXIA</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCE_H vs SCE_L</td>
<td>1.8</td>
<td>0.031</td>
</tr>
<tr>
<td>SCE_H vs SCE_M</td>
<td>1.41</td>
<td>0.129</td>
</tr>
<tr>
<td>SCE_M vs SCE_L</td>
<td>1.72</td>
<td>0.034</td>
</tr>
<tr>
<td><strong>HALLMARK_INFLAMMATORY_RESPONSE</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCE_H vs SCE_L</td>
<td>2.01</td>
<td>0.018</td>
</tr>
<tr>
<td>SCE_H vs SCE_M</td>
<td>1.81</td>
<td>0.019</td>
</tr>
<tr>
<td>SCE_M vs SCE_L</td>
<td>1.89</td>
<td>0.015</td>
</tr>
<tr>
<td><strong>HALLMARK_IL6_JAK_STAT3_SIGNALING</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCE_H vs SCE_L</td>
<td>1.84</td>
<td>0.034</td>
</tr>
<tr>
<td>SCE_H vs SCE_M</td>
<td>1.61</td>
<td>0.069</td>
</tr>
<tr>
<td>SCE_M vs SCE_L</td>
<td>1.71</td>
<td>0.033</td>
</tr>
<tr>
<td><strong>HALLMARK_ANGIOGENESIS</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCE_H vs SCE_L</td>
<td>1.76</td>
<td>0.037</td>
</tr>
<tr>
<td>SCE_H vs SCE_M</td>
<td>1.98</td>
<td>0.004</td>
</tr>
<tr>
<td>SCE_M vs SCE_L</td>
<td>1.33</td>
<td>0.182</td>
</tr>
<tr>
<td><strong>HALLMARK_KRAS_SIGNALING_UP</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Key gene networks identified in BLCA stem cell subtypes**

A total of 14 gene modules were generated based on WGCNA. Among them, brown and black modules showed the strongest relationships with stem cell subtypes, and the brown module was positively correlated with SCE_L and negatively correlated with SCE_H and SCE_M. The black module was positively correlated with SCE_H and negatively correlated with SCE_M and SCE_L (Figure 10A). Two pairs of transcription factor regulators were identified in the brown and black modules (Figure 10B–C), IRF5 and GATA3 and GRHL2 and GATA6. IRF5 is a key transcription factor regulating the differentiation of M1 macrophages into M2 macrophages, enabling M2 macrophages to play an anti-inflammatory role; these cells also have an important role in tissue repair and reconstruction and cancer occurrence [33-35]. GATA3 is type 2 helper T cell (Th2) cytokine-specific transcription factor and a key stemness gene that regulates cell differentiation. It enables Th2 to express IL-4 and other cytokines, promotes antibody production, mediates humoral immunity, and suppresses anti-tumor immunity [36-38]. Therefore, GATA3 may act as a tumor suppressor gene in BLCA. GRHL2 and GATA6 play important regulatory roles in many life activities such as embryonic development, damage repair, epidermal barrier formation, tracheal epithelial formation, and neural tube development, but they also play an important role in the occurrence and development of tumors, cell proliferation, invasion, and metastasis are closely related, which were verified in the relevant pathways identified in the black module [39-45] (Figure 10D–E). Thus, we identified GATA3 and GATA6 as important transcription factors with opposite expression and effects on prognosis in patients with different BLCA stem cell subtypes (Figure 10F–I).

**Discussion**

BLCA is one of the main malignant tumors that endangers human health. Although many molecular genotyping schemes for BLCA have been proposed, these studies are still in their infancy compared with breast cancer molecular typing. A unified, well-developed, and highly feasible molecular typing scheme is required to guide the diagnosis and treatment of BLCA. Additionally, no studies have classified BLCA based on stem cell gene sets, and thus we used specific stem cell gene sets to identify and verify our new classification for BLCA. Our results indicate that BLCA can be divided into three stable subtypes: SCE_H, SCE_M, and SCE_L. Among them, the SCE_H subtype showed the highest degree of immune infiltration and lowest tumor purity relative to the other subtypes, but these patients showed the worst prognosis.
This appears to contradict the previous suggestions that a higher degree of tumor immune infiltration is associated with better prognosis. Through CIBERSORT analysis of the immune cell fraction of stem cell subtypes, we found that the proportion of various cells associated with cytotoxicity in the SCE_H subtype was significantly lower than in the other subtypes, such as resting NK cells, CD8 T cells, CD4 T cells, etc. The number of macrophages M0, M1, and M2 was significantly increased. The composition of immune cells in the tumor microenvironment is complex and has different roles in various stages of tumor progression. Among them, macrophages show the highest content in tumor tissues and most significant regulatory effect on tumors; these cells can promote the proliferation, invasion, and metastasis of tumor cells and induce tumor cells to develop immune tolerance. These cells are known as tumor-associated macrophages (TAMs) (most studies have suggested that TAMs are mainly the macrophage M2 type) [46]. TAMs are often distributed around CSCs, and the amount of infiltration is closely related to the tumor histological grade and number of CSCs. Jinushi et al. [47, 48] found that the growth factor and inflammatory cytokine MFG-E8 and IL-6 secreted by TAM led to activation of the STAT3 and sonic hedgehog signaling pathways, which induced CSC formation and enhanced CSC tumorigenesis and resistance to chemotherapy. This is consistent with our survival analysis showing high ssGSEA scores for macrophage M0 and M2 and predicting a poor prognosis for patients with BLCA.

In addition, the immune checkpoint molecule CD274 (PD-L1) was significantly up-regulated in the SCE_H subtype. This molecule suppresses the proliferation and differentiation of T lymphocytes, promotes the differentiation of Tregs, and induce the secretion of cytokines, thereby suppressing the immune response [46]. Prediction of the anti-PD-1 treatment response showed that SCE_H is more sensitive to anti-PD-1 than other subtypes, indicating that CD274 highly expressed in tumor/tumor stem cells and may be involved in the immune escape process of the tumor. SCE_H mainly corresponds to the luminal-infiltrated and basal-squamous molecular subtypes of BLCA. Multiple studies have shown that these two types of tumors exhibit high levels of immune infiltration and respond very well to immune checkpoint therapy (PD-1, PD-L1, CTLA4). This demonstrates that the stem cell classification we defined is closely correlated with the existing molecular typing of BLCA.

Next, we explored the biological changes caused by different levels of BLCA stem cell enrichment. The results showed that SCE_H was positively correlated with EMT, hypoxia, angiogenesis, and inflammatory response activation in the tumor microenvironment. In the pathophysiology of tumors, studies have confirmed that early tumor cells are in an epithelioid state, and as the tumor progresses, more mesenchymal features are gradually obtained. Such mesenchymal cells show increased resistance to clinical treatment options. In addition, activation of EMT in tumor cells induces the initial state of the tumor, also known as the CSC state, suggesting that EMT is an integral process in the progression of all types of malignant tumors [49]. In a breast tumor progression model, Morel et al. [50, 51] showed that after activation by the Ras-mitogen-activated protein kinase pathway, EMT induction can drive breast epithelial cells to obtain stem cell and tumorigenic properties of CSC. Additionally, activation of the hypoxic pathway makes cancer cells more adaptable to the hypoxic environment. Under hypoxic conditions, the hypoxia-inducible factor (HIF-1α) pathway is activated to promote the release of vascular endothelial growth factor and platelet-derived growth factor, and these cytokines can cause endothelial
cells from the original tumor blood vessels to proliferate, bud, and generate new tumor blood vessels, so that the tumors invade and metastasize. Notably, immune cells infiltrating the tumor microenvironment can secrete a large number of cytokines and chemokines to promote EMT in tumor cells. Further, uncontrollable inflammatory lesions can regulate the occurrence of EMT in tumor cells, and a positive feedback loop can be formed between the inflammatory lesions and EMT, so that the EMT process and uncontrollable inflammatory state can be maintained. These common pathways constitute a vicious circle of tumorigenesis, development, drug resistance, and even poor prognosis.

**Conclusion**

By identifying BLCA subtypes based on stem cell gene sets, we systematically analyzed the relationship between this subtype in the tumor microenvironment and immune cells, immunotherapy response, corresponding pathways, and key genes. These results provide a basis and reference for the clinical diagnosis and treatment of BLCA.

**Abbreviations**

SCE_H: high stem cell enrichment; SCE_M: medium stem cell enrichment; SCE_L: low stem cell enrichment; BLCA: Bladder cancer; CSC: cancer stem cell; TCGA: The Cancer Genome Atlas; GEO: Gene Expression Omnibus; FPKM: fragments per kilobase of transcript per million mapped reads; ssGSEA: single-sample gene set enrichment analysis; CDF: cumulative distribution function; OS: overall survival; RFS: relapse-free survival; PFS: progression-free survival; CTL: cytotoxic T cell; WGCNA: weighted correlation network analysis; ANOVA: analysis of variance; Tregs: regulatory T cells; OCLR: one-class logistic regression; EMT: epithelial mesenchymal transition; NES: normalized enrichment score; FDR: false discovery rate; Th2: type 2 helper T cell; TAMs: tumor-associated macrophages; HIF-1α: hypoxia-inducible factor; GSEA: gene set enrichment analysis.

**Declarations**

**Ethics approval and consent to participate**

All data was downloaded from public databases and therefore did not require approval and review by the ethics committee.

**Consent for publication**

Not applicable

**Availability of data and materials**

The datasets generated or analyzed for this study can be found in the TCGA Knowledge Base (https://portal.gdc.cancer.gov/repository), GEO (https://www.ncbi.nlm.nih.gov/geo/) and ArrayExpress
(https://www.ebi.ac.uk/arrayexpress/) database. 26 stem cell gene sets were obtained from StemChecker (http://stemchecker.sysbiolab.eu/).

Competing interests

The authors declare that they have no competing interests

Funding

This work was supported in part by the National Natural Science Foundation of China (2016GXNSFAA380306), a self-generated project of the Guangxi Zhuang Autonomous Region Health Department (Z20170816), China.

Authors’ contributions

CT analyzed and interpreted the patient data and was a major contributor in writing the manuscript. JM proposed the conception and performed data analyses. XL was responsible for the critical work of the manuscript. ZL corrected the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Editage (www.editage.cn) for English language editing.

References


**Supplementary Materials**

Supplementary Table 1. The 26 stem gene sets used for identification of BLCA subtype.

Supplementary Figure 1. Evaluation of immune cell infiltration level, tumor purity, and stromal content in BLCA. (A-F) Immune score, (G-L) stromal score (stromal content) and (M-R) tumor purity in all six data sets. *P < 0.05, **P < 0.01, ***P < 0.001, ns means not significance.

Supplementary Figure 2. Comparisons of the expression levels of immune-related genes between BLCA subtypes. (A-C) Expression situations of HLA genes between BLCA subtypes in TCGA, E-MTAB-4321 and GSE32894. (D-E) Expression situations of immune cell subgroup marker genes between BLCA subtypes. Kruskal-Wallis test, *P < 0.05, **P < 0.01, ***P < 0.001, ns means not significance.

Supplementary Table 2. Univariate Cox analysis for all six data sets.

**Figures**
Figure 1

Workflow chart.
Figure 2

Identification of three stem cell subtypes in BLCA TCGA cohort. (A) Consensus score matrix for BLCA samples when k = 3. A higher consensus score between two samples indicates that they are more likely to be assigned to the same cluster in different iterations. (B) Cumulative distribution function (CDF) curve describes a real random variable of its probability distribution based on consensus scores for different subtype numbers (k = 2–9). (C) Delta area curve of all samples when k = 3. (D) Three-dimensional plot for ssGSEA scores in three stem cell subtypes; each dot indicates a sample, and different colors represent different subtypes. (E) Heatmap of ssGSEA scores for three subtypes.
Figure 3

Differential expression of CD274 in BLCA stem cell subtypes. Expression of immune checkpoint CD274 (PD-L1) in (A) TCGA, (B) GSE13507, (C) GSE31684, (D) GSE32548, (E) GSE32894, and (F) E-MTAB-4321 cohort determined by ANOVA.

Figure 4
Survival analysis of patients with different BLCA stem cell subtypes. Comparison of survival prognosis between BLCA subtypes in overall survival (OS) of (A) TCGA, (F) GSE31684, (G) GSE32548, and (H) GSE32894 and relapse-free survival (RFS) in (B) TCGA together with progression-free survival (PFS) in (C) TCGA, (D) E-MTAB-4321, and (E) GSE13507.

**Figure 5**

BLCA stem cell subtypes immunotherapy response prediction. (A) SCE_H vs SCE_L sensitive response to PD1 inhibitor and CTLA4 inhibitor (Benjamini & Hochberg corrected P = 0.036, P = 0.036); (B) SCE_H vs SCE_M sensitive response to PD1 inhibitor and CTLA4 inhibitor (Benjamini & Hochberg corrected P > 0.05, P = 0.008); (C) SCE_M vs SCE_L sensitive response to PD1 inhibitor and CTLA4 inhibitor (Benjamini & Hochberg corrected P = 0.048, P > 0.05); (D) Sankey chart showing the distribution of BLCA stem cell subtypes in C1–C6 (C5 was not available for BLCA) and molecular subtypes.
Figure 6

Difference analysis of 22 human immune cell subgroups of BLCA stem cell subtypes in CIBERSORT. Immune cell subgroups with significant differences in BLCA stem cell subtypes in (A) TCGA; (B) GSE32894; (C) GSE31684; (D) E-MTAB-4321; (E) GSE13507, and (F) GSE32548 cohort with CIBERSORT. Fraction of different immune cell subgroups among the four subtypes evaluated by Kruskal–Wallis tests, * P < 0.05, ** P < 0.01, *** P < 0.001. Kaplan-Meier survival curve based on median ssGSEA score for (G) TCGA; (H) GSE13507; (I) GSE32548; (J) GSE32894, and best cut-off for (K) E-MTAB-4321 cohort in OS for macrophage M0; together with median ssGSEA score for (L) TCGA in OS for macrophage M2.
Figure 7

Stem cell stratification analysis of BLCA based on CD274 and stem genes expression. The scatter plot shows the gene expression value of CD274 and stemness genes (CD44, GATA3, HIF1A, ID1, MYC, SOX9, CXCL8) after log2 conversion (A–G). Log rank test was used to determine the best cut-off value to divide continuous variables of gene expression into high and low expression groups. The blue arrow shows the highest test score of the candidate cut-off point based on the gene expression value of the log2
transformation, so the patients were divided into 4 different risk groups (blue dotted line, groups I–IV). The risk groups all groups with their proportions expressed as percentages. A linear regression line was drawn; the gray shaded area shows the 95% confidence interval, and correlation analysis revealed the Pearson correlation coefficient. The SCE_H (black triangle), SCE_M (gray circle), and SCE_L (brown diamond) status are marked as BLCA samples. (H) Correlation diagram of CD274 and stem genes, Pearson correlation coefficient, red indicates positive correlation, green indicates negative correlation. (I–O) Kaplan-Meier survival curves of risk groups I and III were plotted, log-rank test.

Figure 8
Clinical and subtype characteristics associated with the stemness index based on mRNA expression. (A–F) Box plot of mRNAsi in the 6 datasets, stratified by the BLCA stem cell subtype using our definition, Kruskal-Wallis test. (G–I) Box plots of mRNAsi in TCGA, GSE13507, and GSE32894 according to the comparison of normal and cancer tissue samples, and the box plot of mRNAsi of GSE13507 primary and recurrent BLCA samples; (J) A. Gordon Robertson and (K) Lund University BLCA molecular typing mRNAsi box plot. (L–P) Box plots of BLCA samples classified by T stage in TCGA, GSE31684, GSE32548, GSE32894, and E-MTAB-4321 cohorts. The Wilcoxon test was used between the two groups, and the Kruskal-Wallis test was used for groups of three and more.

Figure 9

Figure 10

Identification of key genes / networks based on WGCNA for BLCA stem cell subtypes. (A) Correlation heat map of 14 modules in 3 stem cell subtypes obtained from 4876 differential genes compared to in normal tissues and BLCA of TCGA cohort. (B) In the brown module, the co-expression network is composed of IRF5 and GATA3; (C) In the black module, the co-expression network is composed of GRHL2 and GATA6; (D) Top 20 enrichment pathways obtained by using Metascape in the black module; (F) Top 20 enrichment channels arranged by -Log10P value in the black module. (G) Differential expression of GATA3 and (H) GATA6 in BLCA stem cell subtypes, Kruskal–Wallis tests, * P < 0.05, ** P < 0.01, *** P <
0.001. (I) GATA3, and (J) GATA6 Kaplan-Meier survival curve according to the median expression level using GEPIA, log rank test.

Supplementary Files

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

- SupplementaryTable1.xlsx
- SupplementaryFigure1.tif
- SupplementaryTable2.xlsx
- SupplementaryTable1.xlsx
- SupplementaryFigure1.tif
- SupplementaryTable2.xlsx
- supplementaryFigure2.tif
- supplementaryFigure2.tif