FOXA1, GATA3 and PPARγ Cooperate to Drive Luminal Subtype in Bladder Cancer: A Molecular Analysis of Established Human Cell Lines

Joshua I. Warrick1,2*, Vonn Walter3,4, Hironobu Yamashita1, Eunah Chung5, Lauren Shuman1,2, Vasty Osei Amponsa1, Zongyu Zheng1, Wilson Chan1,2, Tiffany L. Whitcomb5, Feng Yue3,6, Tejaswi Iyyanki3, Yuka I. Kawasawa3,6, Matthew Kaag2, Wansong Guo7, Jay D. Raman3, Joo-Seop Park5, David J. DeGraff1,2,3*

1Department of Pathology, Pennsylvania State University Milton S. Hershey Medical Center, Hershey PA
2Department of Surgery, Division of Urology, Pennsylvania State University College of Medicine
3Department of Biochemistry and Molecular Biology, Pennsylvania State University College of Medicine
4Department of Public Health Sciences, Pennsylvania State University College of Medicine
5Department of Pediatric Urology and Developmental Biology, Cincinnati Children’s Hospital Medical Center
6Institute for Personalized Medicine, Pennsylvania State University College of Medicine
7Department of Surgery, Division of Urology, Changchun Central Hospital, Changchun, China

*Corresponding Authors

Supplemental protocols

TCGA and CCLE data comparison

For comparison between CCLE cell lines and human tumors, TCGA data from muscle invasive urothelial bladder cancer were obtained from the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/). RNA-seq and mutational data were included. Cases with expression subtyping data were selected (n=127, https://tcga-data.nci.nih.gov/docs/publications/blca_2013). TCGA type 1 and 2 tumors were considered Luminal, and types 3 and 4 considered basal15. Similarly, CCLE expression data by microarray, mutational data (Oncomap), and SNP array data were included (http://www.broadinstitute.org/ccle/home).

Copy number alterations for TCGA data were taken directly from gene-level copy number calls from firebrowse.org (http://firebrowse.org/?cohort=BLCA&download_dialog=true). Based on these data, percentiles of discreet copy number alterations were determined, yielding: homozygous loss (-2), 0.69%; heterozygous loss (-1), 24%; copy number neutral (0), 51%; low-level copy number gain (1), 23%; high-level copy number gain (2), 2%. CCLE .seg files were used to determine copy number alterations in cell lines. Segmentation mean values falling into the approximate percentiles as the TCGA were utilized, yielding: homozygous loss (-2), segmentation mean=-0.4; heterozygous loss (-1), segmentation mean=-0.2; low-level copy number gain (1), segmentation mean=0.2; high-level copy number gain (2), segmentation mean=0.4. Veracity of this approach in the CCLE data was validated by comparing CDKN2A copy number loss assignment with CDKN2A
expression levels, which showed marked loss of expression in cases with homozygous CDKN2A copy number loss (p<0.001, Figure S2).

**Expression clusters in CCLE data**

Normalized microarray-derived expression data for urinary tract cell lines (n=27) were downloaded from the CCLE data portal (http://www.broadinstitute.org/ccle/data/browseData?conversationPropagation=beginc). The list of 2,708 gene used in the TCGA study was downloaded (https://tcga-data.nci.nih.gov/docs/publications/blca_2013/BLCA_gnames.txt), and the CCLE expression matrix was limited to these genes (2,556 genes from list in CCLE bladder cancer expression data). The expression matrix was log2 transformed. Agglomerative clustering was performed using correlation distance on the log2 transformed expression matrix using Ward’s method, McQuitty’s method, complete linkage, and average linkage to produce four dendrograms (figure S3). Ward’s and McQuitty’s methods placed cell lines in the same clusters when each was cut into three separate clusters (figure S2 green lines). Complete linkage and average linkage gave similar results, though these methods placed a small number of cell lines in their own smaller clusters. Expression subtype of cell lines was assigned as luminal, basal, or non-type based on the cluster assignments identified by Ward’s and McQuitty’s methods.

Cluster centroids were determined using cluster assignments described above, using luminal and basal cell CCLE lines only, based on log2 normalized data. The ClaNC package was utilized. Cross-validation errors rates were minimal when including 25 genes per subtype (class), for a total of 50 genes used in centroid construction. A dendrogram (agglomerative clustering, average linkage, Euclidean distance) using these 50 genes showed clearly separated clusters (figure S1A). Correlation distance between each cell line (including our modified cell lines and the CCLE cell lines, both log2 normalized and scaled) and the luminal and basal centroids was determined. Modified cell lines were assigned to the subtype with the greatest correlation to the corresponding centroid vector (Figure 5).

A list of ~1,500 human transcription factors was obtained. Differential expression of these transcription factors was determined between luminal (type I and II) and basal (types III and IV) cancers in the TCGA case set detailed above using the limma package in R, using TCGA data from the 127 cases detailed above. 468 transcription factors were identified as overexpressed in luminal tumors (q<0.1, t-statistic from linear model),
and 194 overexpressed in basal tumors (q<0.1, t-statistic from linear model). Using this selected set of 662 transcription factors, expression levels were compared between our control HTB9 cell lines and matched FOXA1.GATA3.PPARG cell lines. Specifically, transcription factors with fold-2 change in the expected direction (i.e. elevated in luminal genes and decreased in basal genes) were identified in 82 luminal genes and 34 basal transcription factor genes (Figure 6A).

**Supplemental Figure Legends**

ST1: Transcription factor motifs identified as being associated with FOXA1 occupied regions by ChIP-Seq in RT4 bladder cancer cells.

S1: Hierarchical clustering of Luminal and Basal cell lines using the 50 gene list used to define Luminal and Basal centroids, derived from the ClaNK package. This gene list robustly separates Luminal from Basal in both (A) CCLE data and (B) TCGA data.

S2: CDKN2A expression by CDKN2A copy number status in CCLE cases. Markedly low expression is present in cases with homozygous CDKN2A loss vs copy number neutral (p<0.001, Wilcoxon rank sum test).

S3: Hierarchical clustering dendrograms for CCLE data using multiple methods. Green lines indicate where dendrograms were cut for cluster assignment.

S4: Immunohistochemistry of RT4, UMUC1 and SCaBER tissue recombinants for KRT20 and members of the uroplakin family of proteins.

S5: KRT5/6 and EGFR western blotting of cell lysates extracted from 5637 bladder cancer cells transiently transfected with empty vector, FOXA1, GATA3 and FOXA1 and GATA3, and subsequently treated in the presence or absence of rosiglitazone. See materials and methods for description of experimental approach.

S6: Heatmap display of genes that are differentially expressed between treated and normal cell lines following the overexpression of FOXA1 and GATA3 in 5637 bladder cancer cells treated with rosiglitazone (n = 310 up in treatment, n = 730 down in treatment). Genes appear in rows and the expression values are median centered by gene.
S7: MSigDB Pathway analysis of FOXA1 ChIP-Seq data, with a focus on genes associated with FOXA1 occupied regions and associated GATA3 and PPARγ binding sites identifies an association with the PDGFR-alpha signaling pathway.

S8: MSigDB Perturbation analysis of FOXA1 ChIP-Seq data with a focus on genes associated with FOXA1 occupied regions and associated GATA3 and PPARγ binding sites

S9: Analysis of FOXA1 ChIP-Seq data to identify distance of FOXA1 occupied regions with associated GATA3 and PPARγ binding sites to the transcription start site (TSS) to a given gene.
Figure S2
<table>
<thead>
<tr>
<th></th>
<th>RT4</th>
<th>UMUC1</th>
<th>SCaBER</th>
</tr>
</thead>
<tbody>
<tr>
<td>H&amp;E</td>
<td><img src="image1" alt="H&amp;E Image" /></td>
<td><img src="image2" alt="H&amp;E Image" /></td>
<td><img src="image3" alt="H&amp;E Image" /></td>
</tr>
<tr>
<td>KRT20</td>
<td><img src="image4" alt="KRT20 Image" /></td>
<td><img src="image5" alt="KRT20 Image" /></td>
<td><img src="image6" alt="KRT20 Image" /></td>
</tr>
<tr>
<td>UPK1A</td>
<td><img src="image7" alt="UPK1A Image" /></td>
<td><img src="image8" alt="UPK1A Image" /></td>
<td><img src="image9" alt="UPK1A Image" /></td>
</tr>
<tr>
<td>UPK1B</td>
<td><img src="image10" alt="UPK1B Image" /></td>
<td><img src="image11" alt="UPK1B Image" /></td>
<td><img src="image12" alt="UPK1B Image" /></td>
</tr>
<tr>
<td>UPK2</td>
<td><img src="image13" alt="UPK2 Image" /></td>
<td><img src="image14" alt="UPK2 Image" /></td>
<td><img src="image15" alt="UPK2 Image" /></td>
</tr>
<tr>
<td>UPK3A</td>
<td><img src="image16" alt="UPK3A Image" /></td>
<td><img src="image17" alt="UPK3A Image" /></td>
<td><img src="image18" alt="UPK3A Image" /></td>
</tr>
</tbody>
</table>