Interpretative machine learning approach reveals gene expression biomarkers predicting cancer patient outcomes at early stages

**Abstract**

In order to understand the molecular mechanisms underlying early cancer development, we developed an interpretable, data-driven machine learning approach to identify the gene biomarkers that predict the clinical outcomes of early cancer patients. As a demonstration, we applied this approach into large-scale pan-cancer datasets including TCGA[1] to find out how effective it would be at identifying the developmental gene expression biomarkers across tumor stages for various cancer types. More relevant to the goal of machine learning interpretable classifiers, we found that early cancer patient groups clustered by the biomarkers selected have significantly more survival differences than ones by early TMM stages, suggesting that this method identified novel early cancer molecular biomarkers. Furthermore, using lung cancer as a study case, we leveraged the hierarchical architectures of neural network to identify the developmental regulatory networks controlling the expression of early cancer biomarkers, providing mechanistic insights of functional genomics driving the onset of cancer development. Finally, we reported the drugs targeting early cancer biomarkers, revealing potential genomic medicine affecting the early stage cancer development. The resulting computational methods are provided with open source in the public domain.

**Interpretable machine learning framework**

**Hierarchical in artificial neural network reveals biological functions and pathways across cancer stages for lung adenocarcinoma**

**Gene regulatory network controlling gene expression biomarkers predicting cancer stages in lung adenocarcinoma**

**Pan-cancer**

We applied our machine learning approach to multiple cancer types using TCGA datasets, and identified the cancer-type-specific gene expression biomarkers for early stage. We found that the patient groups clustered by these biomarkers (e.g., using Partitioning Around Medoids, Hierarchical Clustering) have significantly different survival rates than the ones grouped by cancer stages (log-rank test).

**Drug Biomarker Gene Stage**

<table>
<thead>
<tr>
<th>Drug</th>
<th>Biomarker Gene</th>
<th>Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crizotinib</td>
<td>ALK</td>
<td>Stage IB-ISA Non-Small Cell Lung Cancer that has been removed by surgery and ALK mutations.</td>
</tr>
<tr>
<td>BI-2536</td>
<td>FAM170A, CYP4B1, VCAM1</td>
<td>NCC</td>
</tr>
<tr>
<td>Navitoclax elatinib Cyclophosphamide</td>
<td>SESN1, ECH2C2, AOC3, EC27A3, TGFBR3</td>
<td>Early stage cancer stage</td>
</tr>
</tbody>
</table>

where Y denotes the drug sensitivity variable, Gt. T and B denote the expression of gene i, the tissue source and the experimental batch respectively, and βs are the regression coefficients.

**References**

2. https://www.cancerrxgene.org/
3. Andrew J. Gentles et al. Integrating Tumor and Stromal Gene Expression Signatures With Clinical Indices for Survival Stratification of Early-Stage Non-Small Cell Lung Cancer. Journal of the National Cancer Institute
*Enrichment analysis: Cytoscape (http://www.cytoscape.org)