TCGA computational histopathology pipeline reveals subtypes and their molecular signature

Hang Chang, Ju Han, Cemal Bilgin, Gerald Fontenay, Alexander Borowski, Joe Gray, Paul Spellman, and Bahram Parvin

Lawrence Berkeley Laboratory
Computational histopathology pipeline captures molecular basis for each morphometric subtype.
Use case and target for analysis

- Glioblastoma multiforme (GBM)
  - Curated by removing tissue sections with artifacts (e.g., fold in tissue, pen mark, scanning anomaly)
  - Sample size
    - 380 tissue sections selected out of 447
    - 146 patients selected out of 152

- Challenges?
  - Technical and biological variations, very large datasets

- Approach
  - Development of robust and efficient image analysis algorithms
  - Computing morphometric features and meta-features
  - Subtyping based on selected features or reduced dimensionality (e.g., PCA, MDS)
  - Molecular association with morphometric subtypes
New algorithm enhances nuclear segmentation in the presence of technical variations
Seed detection provides shape signature and local statistics
Cell-by-cell segmentation result
Cell-by-cell segmentation result
Representation

Structural features

Cell-by-cell measurement

A vector

A multidimensional distribution

Normalization across all tissue sections
What are subtypes based on cellularity and nuclear size at the patient level

2 clusters

3 clusters

4 clusters

5 clusters

6 clusters
What is the distribution of each subtype and how well each subtype predicts survival as a function of treatment?

Subtype 2
What are the molecular basis of each subtype?

- **Gene selection**
  - Univariate or multivariate methods
  - **Pathway** or subnetwork enrichment analysis

<table>
<thead>
<tr>
<th>Name</th>
<th>Overlapping Entities</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focal Adhesion Regulation</td>
<td>CAV1,MET,ERBB4,KIT,PDGFRA,RASA4</td>
<td>0.000208</td>
</tr>
<tr>
<td>Actin Cytoskeleton Regulation</td>
<td>MET,ERBB4,KIT,PDGFRA,SGCE,RASA4,PDLIM3</td>
<td>0.000555</td>
</tr>
<tr>
<td>Gap Junction Regulation</td>
<td>MET,ERBB4,KIT,NPY2R,PDGFRA,RASA4</td>
<td>0.008248</td>
</tr>
<tr>
<td>Adherens Junction Regulation</td>
<td>DAAM2,MET,ERBB4,KIT,PDGFRA,CDH6</td>
<td>0.011068</td>
</tr>
<tr>
<td>KIT -&gt; STAT signaling</td>
<td>KIT</td>
<td>0.017364</td>
</tr>
<tr>
<td>HGFR -&gt; STAT signaling</td>
<td>MET</td>
<td>0.023089</td>
</tr>
<tr>
<td>PDGFR -&gt; STAT signaling</td>
<td>PDGFRA</td>
<td>0.025939</td>
</tr>
<tr>
<td>HGFR -&gt; FOXO3A signaling</td>
<td>MET</td>
<td>0.054015</td>
</tr>
</tbody>
</table>

**Subtype1**

<table>
<thead>
<tr>
<th>Name</th>
<th>Overlapping Entities</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCR1 -&gt; STAT signaling</td>
<td>CCL4,CCL3</td>
<td>0.003127</td>
</tr>
<tr>
<td>CCR5 -&gt; TP53 signaling</td>
<td>CCL4,CCL3</td>
<td>0.004022</td>
</tr>
<tr>
<td>Gap Junction Regulation</td>
<td>GNAO1,CCL4,HRH1,KIT,CCL3,CALCRL,ADCY2,FGF12,RASA4</td>
<td>0.008737</td>
</tr>
<tr>
<td>KIT -&gt; STAT signaling</td>
<td>KIT</td>
<td>0.011068</td>
</tr>
<tr>
<td>HGFR -&gt; FOXO3A signaling</td>
<td>MET</td>
<td>0.017364</td>
</tr>
<tr>
<td>HGFR -&gt; STAT signaling</td>
<td>MET</td>
<td>0.023089</td>
</tr>
<tr>
<td>PDGFR -&gt; STAT signaling</td>
<td>PDGFRA</td>
<td>0.025939</td>
</tr>
<tr>
<td>HGFR -&gt; FOXO3A signaling</td>
<td>MET</td>
<td>0.054015</td>
</tr>
</tbody>
</table>

**Subtype3**

<table>
<thead>
<tr>
<th>Name</th>
<th>Overlapping Entities</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL11R -&gt; STAT3 signaling</td>
<td>IL11RA</td>
<td>0.018322</td>
</tr>
<tr>
<td>ThromboxaneR -&gt; CREB signaling</td>
<td>RASGRP1,GNG4</td>
<td>0.026307</td>
</tr>
<tr>
<td>EphrinR -&gt; actin signaling</td>
<td>EFNB3,SGCE,EPB41L2</td>
<td>0.02702</td>
</tr>
</tbody>
</table>

**Subtype4**
Can tumor composition be characterized?

- Since tumor is heterogeneous, can we query for subtypes at the block levels and learn about tumor composition?
What are the tumor histology subtypes?

- Subtype 1
- Subtype 2
- Subtype 3
- Subtype 4
Does heterogeneity play a role in survival as a result of a more intense therapy?

Loosely defined semantics of high and low!

Low cellularity
High heterogeneity

High cellularity
Low heterogeneity
Another view: Are cellularity and nuclear size correlated? And outcome?

High cellularity and low nuclear size are better predictive of a more aggressive therapy.
Conclusion

- There are many ways to slice through the data and metadata
  - Cellularity, nuclear size
  - Heterogeneity
- Different indices lead to alternative subtypings
  - Alternative biological interpretation is possible
- Genomic association has the potential to reveal new insight
- Web site: tcga.lbl.gov
  - “Google map” like viewing of tissue sections with segmentation results overlaid