**Model**

**Dataset** includes full base pair sequence of one individual leukemia patient.

- **23** chromosomes sequenced at the base pair level.
- **Chromatin accessibility** of 18 cell types:

**Problem**

Leukemia results from imbalance in regulation of the typical protein pathways.

The 18 related cell types are shown above:

Gene therapy is possible if genetic sequences could be associated with certain diseases and traits (Hindorff et al.). Specifically, we want to map genetic sequences to a chromatin accessibility binary, which can be used to identify anomalies in the protein pathways of these 18 related cell types.

**Solution:**

We achieve our test-metric goal of (0.70 auPRC) and accuracy of (0.90) using a CNN model. Now that we have predicted chromatin accessibility binaries, we can interrogate the model and interpret results.

We identify genetic motifs that strongly activate our first convolutional layer filters using only 6 first layer filters:

Increasing performance by using all 50 first layer filters, we find that we cannot accurately distinguish regulatory genes beyond the control group.

**Confusion Matrix:**

<table>
<thead>
<tr>
<th>Label</th>
<th>Predict: 0</th>
<th>Predict: 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.46</td>
<td>0.4</td>
</tr>
<tr>
<td>1</td>
<td>0.54</td>
<td>0.6</td>
</tr>
</tbody>
</table>

**References**