**Lung Cancer Type Predictor: Using Spark**

Group 16  
Xander Krohannon

Gayathri Panangipalli

Sharmila Selvaraj

**Problem statement:**

Precision medicine; the breakthrough in science in the late 1990’s, seems to be well on its way to becoming a significant approach to treatment of diseases. In the relatively short time since its advent, research has moved toward the discovery of monoclonal antibodies and small molecule inhibitors as potential targeted treatments for cancer. Although, these drugs act directly on cancerous cells, multiple cases of off-target effects on normal cells have also been reported. With the significant increase in the number of cancer causative factors, both intrinsic and extrinsic, there exists a gap in knowledge in terms of the genetic pre-disposition for each tissue or type of cancer. This gives room for exploring and defining complex networks of regulation and gene function and dissecting it to design better drugs, tailor-made to each type of cancer. When choosing a target for any drug, one needs to identify either genes which are mutated in cancer cells or genes whose expression is only essential to cancer cells.

Lung cancer, being the second most diagnosed cancer type in the developed countries such as the United States is divided into two major types: Small cell lung cancer (SCLC) and Non-small cell lung cancer (NSCLC). Adenocarcinoma, Squamous cell carcinoma and large cell carcinoma are the three subtypes of NSCLC accounting for 85% of lung cancer incidence. SCLC accounts for 10 – 15% of the total lung cancer incidence but it is one of the very aggressive forms of cancer due to rapid development.

Using the machine learning component of spark, we propose to build a model that can, based on expression levels of genes, differentiate between cancerous tissue and types and provide a consensus that can help frame formulations of more specific drugs and targets.

**Data:**

For this study, we will be collecting transcriptomic data for healthy and lung cancer patients from the following databases:

1. Gene Expression Omnibus (GEO)
2. Gene-tissue expression (GTEx) : expression data for 515 samples
3. The Cancer Genome Atlas (TCGA) : transcriptome data for 1135 cases

**Methodology:**

The RNA-seq data will be downloaded from various sources and compiled together, to expand the number of samples. Afterwards, the numerous reads from each sample will be analyzed using a suite of next generation sequencing tools, including: fastqc, fastx, hisat, and macs. Utilizing those results, expression profiles for control (non-cancerous cells) and each type of cancer will be established.

Finally, the expression data will be fed into a machine learning classifier. A variety of different machine learning model architectures will be used, to determine which gives the best performance. In addition, to determining which model architecture gives the best results, a level of feature refinement will likely be necessary. If this proves necessary, features will be selected based upon either statistical correlation, with a correction factor, or by another calculation: such as mutual information, or GINI significance.

**Outcome:**

From the project, we expect to have a model that, given an expression profile of a specific lung cancer can accurately determine the type of that cancer. In a clinical setting, this could be broadly applied to help pair specific cancer types with targeted therapies, in consortion with traditional approaches: chemotherapy, radiotherapy or more generalized approaches.