Analysis Plan – DNA Methylation in Cancer

Nick Perry & Malaika MB

Intro to Data Science – Fall 2017

1. **Partitioning Data**

To create test and training datasets, the dplyr functions group\_by(Cancer\_Type) and top\_n(x, Cancer\_Type) will be used to subset the master data-frame data from each cancer type into halves (where the value x is positive for test cases, and negative for training cases). The training and testing datasets will contain an equal number of entries, and later provide a means to evaluate our models.

1. **Preparing Visualizations**

A bar graph will be created to visualize the differences in methylation positions between the four types of cancer, faceted by other contributory categorical variables such as feature type and gene type. A boxplot could also be created, showing how each form of cancer has some variation in their own positions of methylation, relative to the transcription start signal. Finally, a depiction of the decision tree could be provided to show the predictive ability between each of the variables. (Perhaps a cumulative frequency plot/joyplot representing methylation position within a gene or chromosome, depending on data coercion and package functionality).

**3. Types of Analyses to be Performed**

**-***Statistical modelling for predictive analysis of new data of unknown cancer type.*

A decision tree will be created to show how methylation position and affected gene type leads to different forms of cancer. A k-nearest neighbor could also be created to potentially improve the predictive power of our model. Confusion matrices will be employed to validate the performance of the trained model on the testing data. Once a well-fit model has been attained, an impartial party will select a single methylation observation (of one of the cancer types explored in this project) unknown to the team, at which time the model will be employed to predict the identity of the unknown cancer type.