1. What features of the data are most important for QC/QA?
2. Missing data are important to be considered. In this case, drop the columns full of nan for both mRNA expression and mutation sequencing data as well as individuals with labels containing missing data (e.g. Neoplasm Histologic Grade and Patient's Vital Status).
3. Since mRNA expression, mutation sequencing and clinical data would be combined together for modeling, it is important to make sure sample IDs were consistent between each other. Here I noticed only 482 samples existed in all three types of data, so in this study only intersection part among three were selected for further modeling in order to combine all features together.
4. Some outliers could be potential errors during sequencing or sampling. Here I utilized PCA to reduce dimensionalities for visualization, followed by checking outliers in top components.
5. Generally speaking, what are potential sources of ambiguity arising from your approach?

It could be from lack of environmental data. As we know, phenotype = clinical data + transcriptome + genotype + environmental data. The environmental data could include information of individuals’ diets, life styles, habits and so on, which would have potential affect on causation of cancers. This could be the biggest sources of ambiguity in our approach.

1. What other data might we collect to enhance risk quantification? What quantitative proof do you have?

If we do whole genome sequencing, whole transcriptome and collected more samples of patients, I think it will be greatly helpful to enhance risk quantification.

For example, we have imbalanced classes for Neoplasm Histologic Grade as showed below. Most of patients were diagnosed at G2 and G3. Assume we had a native classifier that just randomly predict G2 and G3, it will still achieve an about 80% accuracy. This is pretty common to observed in cancer detection, since many people at late grade were already dead. So increasing samples size will definitely increase statistical power and freedom, leading to achieving a better performance of model. Another quantitative approach is to use rarefaction curve that plotted sample size against train and test accuracies/errors. If train and test curves were very different and not converged together, adding more samples will help to reduce variance. If train and test curves were converged, adding samples will not be help to improve this model but it could still be helpful if we switched to a more complex model.

Neoplasm Histologic Grade

G1 14

G2 227

G3 205

G4 76

GX 5

1. Describe your approach to filing IP claims around your unique classification of risk?

Here the random forest was used to train the dataset, achieving an 0.36, 0.5, 0.69 accuracies.

1. How would you communicate your findings to a clinician?

The model we built had a good performance on Disease-associated risk. The model was trained based on genome, transcription and clinical data, which was directly derived from urine and noninvasive. The prediction result could be an additional support for doctors’ diagnose. Despite prediction score is not perfect, we have confidence to achieve a better performance if we can collect more patience data. The more data (both examples and useful features) will help to find similarities and differences within and between classes. So we are looking forwards to more opportunities to collaborate with you!