**Prostate Cancer Dataset**

The problem consists of finding meaningful biomarkers in prostate cancer. This can be done via classification and feature selection for selecting genes that contribute to one or more different classifications. A dataset of 494 samples downloaded from the Genomic Data Commons (formerly cBioPortal) contains gene expressions for a few dozen thousand genes. You are free to work on one or more problems as discussed in class: classification, solving the multi-class problem, feature selection, other aspects, or a combination of these, by using one or more clinical variables (e.g., clinical stage of progression, primary site, Gleason score, etc.).

prad\_tcga\_clinical\_data.xlsx contains the clinical variables for all 494 samples (patients) in the dataset.

prad\_tcga\_genes.xlsx contains gene expression ratios for a few dozen thousand genes.

More details of the clinical variables and the gene expression tables can be found at <http://www.cbioportal.org/data_sets.jsp> . Once there, search for “prostate adenocarcinoma” and choose “Prostate Adenocarcinoma (TCGA, Provisional)” that contains 499 samples.

# Questions

* Why doesn’t the data in the 2 excel files match?
  + Patient TCGA-2A-A8VO has a gleason score of 6 in the clinical data an a gleason score of 9 in the genes data
  + Number of patients is inconsistent (494 in genes data, 499 in the clinical data).
* We can basically combine the two spreadsheets (except they have discrepancies). Are we to use only the gene expression levels as the independent variables / features, or can we use clinical elements as well(example: ethnicity).
* Is there a “right” choice for the target variable, or is this left up to us? It seems as though all of these people have cancer, so obviously we aren’t “predicting” whether or not they will have cancer

### Genes CSV

* What are the row headers – are these code for various genes (and the values are gene expression levels)?

### Clinical CSV